

UNIVERSITY OF SÃO PAULO  
SÃO CARLOS SCHOOL OF ENGINEERING  
RIBEIRÃO PRETO MEDICAL SCHOOL  
SÃO CARLOS INSTITUTE OF CHEMISTRY

FILIPE ANDRADE BERNARDI

Digital health research governance: from FAIR to RE-AIM

São Carlos

2024



FILIPPE ANDRADE BERNARDI

Digital health research governance: from FAIR to RE-AIM

Thesis presented to the São Carlos School of Engineering, Ribeirão Preto Medical School and São Carlos Institute of Chemistry of the University of São Paulo in partial fulfillment of the requirements for the degree of Doctor of Science in interunit postgraduate program in bioengineering.

Subject area: Bioengineering

Advisor: Prof. Domingos Alves.

Co-advisor: Prof. Rui Pedro Charters Lopes Rijo.

ORIGINAL VERSION

São Carlos

2024

I AUTHORIZE THE TOTAL OR PARTIAL REPRODUCTION OF THIS WORK,  
THROUGH ANY CONVENTIONAL OR ELECTRONIC MEANS, FOR STUDY AND  
RESEARCH PURPOSES, SINCE THE SOURCE IS CITED.

Catalog card prepared by Patron Service at "Prof. Dr. Sergio  
Rodrigues Fontes" Library at EESC/USP

B523d Bernardi, Filipe Andrade  
Digital health research governance : from FAIR to RE-  
AIM / Filipe Andrade Bernardi; Dissertation directed by  
Domingos Alves; co-advisor Rui Pedro Charters Lopes  
Rijo. -- São Carlos, 2024.  
  
Doctoral (Dissertation) - Graduate Program in  
Bioengineering and Research Area in Bioengineering -- São  
Carlos School of Engineering; Ribeirao Preto Medical  
School; Sao Carlos Institute of Chemistry of the  
University of São Paulo, 2024.  
  
1. Digital health. 2. Governance.  
3. Health research. 4. Data quality. I. Title.

Elena Luzia Palloni Gonçalves – CRB 8/4464



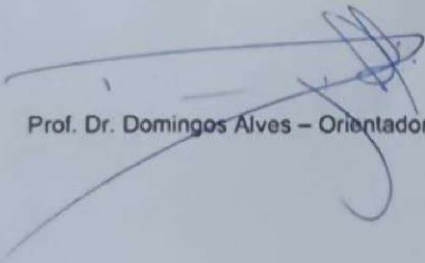
## FOLHA DE JULGAMENTO

Candidato(a): **Filipe Andrade Bernardi**

TÍTULO: "Governança de pesquisas em saúde digital: do FAIR ao RE-AIM"

Data da defesa: 19/04/2024

Comissão Julgadora	Resultado
Prof(a). Dr(a). Domingos Alves Faculdade de Medicina de Ribeirão Preto - FMRP/USP - - Orientador	<u>Não votante</u>
Prof(a). Dr(a). Victor Evangelista de Faria Ferraz Faculdade de Medicina de Ribeirão Preto - FMRP/USP	<u>Aprovado</u>
Prof(a). Dr(a). Mauro Niskier Sanches Universidade de Brasília - UnB	<u>Aprovado</u>
Prof(a). Dr(a). Afrânio Lineu Kritski Universidade Federal do Rio de Janeiro - UFRJ	<u>Aprovado</u>

  
Prof. Dr. Domingos Alves – Orientador





## DEDICATION

Dedicated to my parents, Aerton and Cláudia, I express my profound pride for the education you have provided me and for your constant support in my academic journey and decisions. To my sisters, Rafaela, Fernanda, and Manuela, I am immensely grateful for the emotional and affectionate support, for your dedication and care for our family, and for understanding the absences required by my studies. To all of you, my unconditional love; without the strength of this family pillar, this work would never have been possible.





## ACKNOWLEDGMENTS

To my beloved wife Mariane, your love, patience and unconditional support that you gave me during the long time I dedicated myself to this work were essential.

To my uncle José Cássio, for the encouragement, teachings, and understanding. Thank you very much for your constant support in big and small things

To my great friend Vinicius Lima, whose friendship and partnership were fundamental in the progress of this doctoral thesis. Your encouragement and trust are gifts that I hold gratefully in my memories.

To my dear co-supervisor, Prof. Rui Rijo, for encouraging me and giving me time for the personal construction of the work. The availability he always expressed and the empathy with which he received my ideas were the stimulus that allowed me to overcome the insecurities of this process.

To my advisor, Prof. Domingos Alves, for the opportunity and for believing in your students' potential and sharing this humanized knowledge. I thank you for your trust in my work, creative freedom, respect, understanding and wise life advice.

To the Health Intelligence Laboratory (LIS) for all the infrastructure and support. I thank the team members, especially Diego Yamada and Felipe Pellison, and all researchers and students who shared your collaboration, patience, and dedicated time.

To the professionals of the Brazilian Network for Research on Tuberculosis (REDE-TB) and the National Network for Rare Diseases (RARAS), for the tips, teachings and exchange of experiences that helped guide the activities developed.

To the Interunit Postgraduate Program in Bioengineering - EESC-USP, FMRP-USP, IQSC-USP, director Adair Roberto Aguiar and Marcia Maria Hyppolito Geromini, program secretary.

I thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) for the support through a scholarship during a work period. This study was financed in part by the CAPES – Grant Finance Code 001.

To Fundação de Amparo à Pesquisa do Estado de São Paulo – Brasil (FAPESP - 2020/01975-9) for all the support given to carry out this work

To the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq), for the award of the Technological and Industrial Development (DTI) grant, Level B, which allowed the execution of activities related to this research project.



## EPIGRAPH

“O passado dá-nos responsabilidade. trabalhamos no presente para que o futuro seja melhor e vitorioso... É importante, que em especial os líderes, saibam o motivo de nossas decisões, mesmo que não concordem com elas. Explicar-lhes o porquê das decisões é um passo importante para os convencer e tê-los do nosso lado. Além disso e por vezes, devemos fazer aquilo que é preciso ser feito, e não aquilo que queremos fazer.”

**Abel Ferreira** (Cabeça Fria, Coração Quente - 2022)



## ABSTRACT

BERNARDI, F.A. **Digital health research governance: from FAIR to RE-AIM.** 2024. Thesis (Doctor in Science) - São Carlos School of Engineering; Ribeirao Preto Medical School; São Carlos Institute of Chemistry, University of São Paulo, São Carlos, 2024.

This doctoral research aims to develop a governance model for digital health research data to enhance the quality of collected data and translate research outcomes into an integrated framework for planning, implementing, and evaluating public health initiatives. Employing a mixed-methods approach, the study combines qualitative and quantitative methods, qualifying as an exploratory and explanatory study. The specific objectives include proposing the adoption of frameworks and tools to promote a data quality model for scientific research in digital health, developing a tool with data collection, management, and evaluation, and proposing an implementation research model based on digital health principles and guidelines. The work is grounded in FAIR principles and the RE-AIM framework, addressing data quality and governance in health research, focusing on tuberculosis (TB) and rare diseases (RD) in the Brazilian context. The research significantly contributes to the understanding of digital data governance in health, highlighting the importance of standardized practices and collaborative efforts to improve the quality of health research data in Brazil. Challenges addressed include integrating diverse data sources, ensuring security and confidentiality of data, while enhancing public health outcomes through digital innovations. The presented model emphasizes data quality optimization, being fundamental for planning, executing, and evaluating health interventions. The research operates under robust computational research structures and health information systems, ensuring transparent data management. The interdisciplinary nature of the research highlights its significant strength, integrating insights from data science, public health, and health policy, offering a holistic view of the challenges and opportunities in digital health. By creating a unified implementation quality manual, the study provides mechanisms for network collaboration, strengthening robust data management, and promoting a culture of shared practices essential for elevating the overall quality of data in health research.

Keywords: Digital health. Governance. Health research. Data quality.



## RESUMO

BERNARDI, F.A. **Governança de pesquisas em saúde digital**: do FAIR ao RE-AIM. 2024. Tese (Doutorado) – Escola de Engenharia de São Carlos; Faculdade de Medicina de Ribeirão Preto; Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, 2024.

Esta pesquisa de doutorado tem como objetivo o desenvolvimento de um modelo de governança para dados de pesquisa em saúde digital, visando aprimorar a qualidade dos dados coletados e a tradução dos resultados de pesquisa em um quadro integrado para o planejamento, implementação e avaliação de iniciativas de saúde pública. Utilizando uma abordagem de métodos mistos, a pesquisa combina métodos qualitativos e quantitativos, enquadrando-se como um estudo exploratório e explicativo. Os objetivos específicos incluem propor a adoção de frameworks e ferramentas para promover um modelo de qualidade de dados para pesquisa científica em saúde digital; desenvolver uma ferramenta com diretrizes para coleta, gestão e avaliação de dados; e propor um modelo de pesquisa de implementação baseado em princípios e diretrizes de saúde digital. O trabalho fundamenta-se nos princípios FAIR e no framework RE-AIM, abordando a qualidade de dados e governança na pesquisa em saúde, com foco na tuberculose (TB) e doenças raras (DR) no contexto brasileiro. A pesquisa contribui significativamente para a compreensão de governança de dados digitais na saúde, destacando a importância de práticas padronizadas e esforços colaborativos para aprimorar a qualidade dos dados de pesquisa em saúde no Brasil. Entre os desafios abordados estão a integração de fontes de dados diversas, segurança e manutenção da confidencialidade de dados, ao passo que aprimora resultados de saúde pública através de inovações digitais. O modelo apresentado enfatiza a otimização da qualidade dos dados, sendo fundamental para planejar, executar e avaliar intervenções em saúde. A pesquisa opera sob estruturas computacionais robustas de pesquisa e sistemas de informação em saúde, assegurando a gestão transparente de dados. A natureza interdisciplinar da pesquisa destaca sua força significativa, integrando insights de ciência de dados, saúde pública e política de saúde, oferecendo uma visão holística dos desafios e oportunidades em saúde digital. A partir da criação de um manual unificado de qualidade de implementação, o estudo oferece mecanismos para colaboração em rede, o fortalecimento da gestão de dados robustos e a promoção de uma cultura de práticas compartilhadas essenciais para elevar a qualidade geral dos dados na pesquisa em saúde.

Palavras-chave: Saúde digital. Governança. Pesquisas em saúde. Qualidade de dados.





## LIST OF PUBLICATIONS

- I. **BERNARDI, F. A.**; ALVES, D.; CREPALDI, N.; YAMADA, D. B.; LIMA, V. C.; RIJO, R. Data quality in health research: integrative literature review. *Journal of Medical Internet Research*, v.25, Oct. 2023. DOI: [10.2196/41446](https://doi.org/10.2196/41446).
- II. **BERNARDI, F. A.**; OLIVEIRA, B. M.; YAMADA, D. B.; ARTIFON, M.; SCHMIDT, A. M.; SCHEIBE, V. M; ALVES, D.; FÉLIX, T. M. The Minimum data set for rare diseases: systematic review. *Journal of Medical Internet Research*, v.25, July 2023. DOI: [10.2196/44641](https://doi.org/10.2196/44641).
- III. ALVES, D.; YAMADA, D. B.; **BERNARDI, F. A.**; CARVALHO, I.; COLOMBO FILHO, M. E.; NEIVA, M. B.; LIMA, V. C.; FÉLIX, T. M. Mapping, infrastructure, and data analysis for the Brazilian network of rare diseases: protocol for the RARASnet observational cohort study. *JMIR Research Protocols*, v. 10, n. 1, Jan. 2021. DOI: [10.2196/24826](https://doi.org/10.2196/24826).
- IV. **BERNARDI, F. A.**; ALVES, D.; NEIVA, M. B.; YAMADA, D. B.; LIMA, V. C.; VINCI, A.; THOMAZINI, G.; RIJO, R.; FELIX, T. M. A Proposal for a set of attributes relevant for Web portal data quality: the brazilian rare disease network case. *Procedia Computer Science*, v. 219, p. 1316-1324, 2023. DOI: [10.1016/j.procs.2023.01.416](https://doi.org/10.1016/j.procs.2023.01.416).
- V. **BERNARDI, F.**; LIMA, V.; SARTORETTO, G.; BAIOCHI, J.; CASSÃO, V.; KRITSKI, A.; RIJO, R.; ALVES, D. From raw data to FAIR data: the FAIRification workflow for brazilian tuberculosis research. *Studies in Health Technology and Informatics*, v. 29, n. 305, p. 331-334, June 2023. DOI: [10.3233/SHTI230497](https://doi.org/10.3233/SHTI230497).
- VI. MOZINI, M.; **BERNARDI, F.**; MIOTO, A. C.; CASSÃO, V.; KRITSKI, A.; ALVES, D. A Computational infrastructure for analyzing tuberculosis research data in Brazil. *Studies in Health Technology and Informatics*, v. 305, p. 558-561, June 2023. DOI: [10.3233/SHTI230557](https://doi.org/10.3233/SHTI230557).
- VII. TEIXEIRA, J. C. C.; **BERNARDI, F. A.**; RIJO, R. P. C. L.; ALVES, D. Proposal for a health information management model based on lean thinking. *Procedia Computer Science*, v. 181, n. 5, p. 1097-1104, Jan. 2021. DOI: [10.1016/j.procs.2021.01.306](https://doi.org/10.1016/j.procs.2021.01.306).
- VIII. CREPALDI, N. Y.; LIMA, V. C.; **BERNARDI, F. A.**; ALVES, D. An Information system for monitoring tuberculosis cases: implementation research protocol using RE-AIM for a health region in Brazil. *Procedia Computer Science*, v.219, p. 1128-1135, Jan. 2023. DOI: [10.1016/j.procs.2023.01.393](https://doi.org/10.1016/j.procs.2023.01.393).



## LIST OF FIGURES

Figure 1 - PhD Journey .....	34
Figure 2 - The relationships between FAIR activities and RE-AIM. ....	54
Figure 3 - The Activities between FAIR activities and RE-AIM stages.....	55
Figure 4 - PRISMA flowchart with the results of study selection. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses. ....	77
Figure 5 - Elements involved in the research data quality process. Elements involved in the data quality process. FAIR: Findability, Accessibility, Interoperability, and Reuse; ICHGCP: International Conference on Harmonization—Good Clinical Practice; ISO: International Organization for Standardization. ....	88
Figure 6 - PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. MDS: minimum data set. ....	107
Figure 7 - Framework for quality management of rare disease registries. IT: information technology. ....	136
Figure 8 - Main activities of RARASnet.....	140
Figure 9 - Validation and Implementation process of the RARAS Portal.....	161



## LIST OF TABLES

Table 1 – Inclusion and exclusion criteria for eligibility of studies.....	72
Table 2 – Search procedure on databases. ....	73
Table 3 – Search procedure used on databases. ....	75
Table 4 – Improvement strategies for data quality. ....	79
Table 5 – Influencing factors for data quality.....	80
Table 6 – Distribution of quality dimensions in health research. ....	82
Table 7 – Barriers to health data quality. ....	85
Table 8 – Search strategies for each of the selected databases.....	103
Table 9 – General characteristics of selected studies.....	108
Table 10 – Brazilian Tuberculosis Information Systems (BTIS) compliance with the FAIR principles. ....	166
Table 11 – RE-AIM dimensions, study questions and data source.....	188
Table 12 – Doctoral Dissertation Sharing and Implementation Resources .....	196
Table 13 – Mentorship and Academic Development Resources.....	199
Table 14 – Articles outside the theme thesis.....	201
Table 15 – Conference Papers.....	205
Table 16 – Summary of main contributions and recommendations.....	211



## LIST OF ABBREVIATIONS

AI – Artificial Intelligence

CER – Specialized Rehabilitation Center

CINAHL – Cumulative Index of Nursing and Allied Health Literature

CNPq – Brazilian Council for Scientific and Technological Development

CRF – Case Report Forms

DevOps – Development and Operations

EHDI – Early Hearing Detection and Intervention

ERN – European Reference Network

HPO – Human Phenotype Ontology

HIS – Health Information System

ICD – International Classification of Diseases

IEEE – Institute of Electrical and Electronics Engineers

IT – Information Technology

LILACS – Latin American and Caribbean Health Sciences Literature

MDS – Minimum Data Set

NIH – National Institutes of Health

OMIM – Online Mendelian Inheritance in Man

ORDO – Orphanet Rare Disease Ontology

PCC – Population, Concept, and Context

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RARAS – Brazilian National Network of Rare Diseases (Rede Nacional de Doenças Raras)

RD – Rare Disease

RE-AIM – Reach, Effectiveness, Adoption, Implementation, and Maintenance

REDCap – Research Electronic Data Capture

SQL – Structured Query Language

SRDR – Reference Services in Rare Diseases

SRTN – Reference Services for Neonatal Screening

STROBE – Strengthening the Reporting of Observational Studies in Epidemiology

SUS – Unified Health System

WHO – World Health Organization





# TABLE OF CONTENTS

<b>1 INTRODUCTION</b> .....	<b>31</b>
<b>1.1 MOTIVATION</b> .....	<b>35</b>
<b>1.2 OBJECTIVES</b> .....	<b>37</b>
1.2.1 <i>Specific objectives</i> .....	37
<b>1.3 RESEARCH METHODOLOGY</b> .....	<b>38</b>
<b>2 CHAPTERS RELATIONSHIP</b> .....	<b>40</b>
<b>3 THEORETICAL FOUNDATIONS</b> .....	<b>42</b>
<b>3.1 DIGITAL HEALTH</b> .....	<b>42</b>
3.1.1 <i>Digital Health on the World</i> .....	42
3.1.2 <i>Digital Health in Brazil</i> .....	46
<b>3.2 FAIR PRINCIPLES</b> .....	<b>50</b>
<b>3.3 RE-AIM FRAMEWORK</b> .....	<b>51</b>
<b>3.4 CONCEPTUAL MODEL</b> .....	<b>52</b>
3.4.1 <i>Conception</i> .....	52
3.4.2 <i>Application Scenarios</i> .....	55
3.4.3 <i>Process Component</i> .....	59
3.4.4 <i>Development Component</i> .....	60
3.4.5 <i>Electronic Data Capture Systems (EDCs)</i> .....	61
3.4.6 <i>Semantic Component and Standards</i> .....	62
3.4.7 <i>Management and Monitoring Component</i> .....	63
<b>4 DATA QUALITY IN HEALTH RESEARCH: INTEGRATIVE LITERATURE REVIEW</b> .....	<b>67</b>
<b>4.1 INTRODUCTION</b> .....	<b>67</b>
<b>4.2 METHODS</b> .....	<b>69</b>
4.2.1 <i>Study Design</i> .....	69
4.2.2 <i>Health Research</i> .....	70
4.2.3 <i>Types of Approaches</i> .....	70
4.2.4 <i>Types of Interventions and Evaluated Results</i> .....	71
4.2.5 <i>Eligibility Criteria</i> .....	71
4.2.6 <i>Databases and Search Strategies</i> .....	72
4.2.7 <i>Data Collection</i> .....	74
4.2.8 <i>Data Extraction</i> .....	74
4.2.9 <i>Result Presentation</i> .....	75
4.2.10 <i>Data Synthesis</i> .....	75
<b>4.3 RESULTS</b> .....	<b>76</b>
4.3.1 <i>Study Characteristics</i> .....	76
4.3.2 <i>Data Quality Issues and Challenges</i> .....	81
4.3.3 <i>Factors Affecting Data Quality</i> .....	82
4.3.4 <i>Strategies for Improving Data Quality</i> .....	83
4.3.5 <i>Synthesis of Findings</i> .....	84
<b>4.4 DISCUSSION</b> .....	<b>89</b>
4.4.1 <i>Principal Findings</i> .....	89
4.4.2 <i>Comparison With Prior Work</i> .....	91

4.4.3 Strengths.....	93
4.4.4 Limitations .....	93
4.4.5 Future Directions .....	94
4.4.6 Conclusion .....	95
4.4.7 References.....	95
<b>5 THE MINIMUM DATA SET FOR RARE DISEASES: SYSTEMATIC REVIEW.....</b>	<b>99</b>
<b>5.1 INTRODUCTION .....</b>	<b>99</b>
5.1.1 Background .....	99
5.1.2 Objectives.....	101
<b>5.2 METHODS .....</b>	<b>101</b>
5.2.1 Research Question Definition .....	101
5.2.2 Inclusion Criteria.....	102
5.2.3 Exclusion Criteria .....	102
5.2.4 Search Strategy for Selection of Studies.....	102
5.2.5 Data Collection Processes .....	104
5.2.6 Synthesis Methods.....	106
<b>5.3 RESULTS .....</b>	<b>107</b>
5.3.1 Description of Selected Studies.....	107
5.3.2 Domains of Health.....	109
5.3.3 Description of MDSs in the Selected Studies .....	110
5.3.4 Risk of Bias in Included Studies.....	110
<b>5.4 DISCUSSION .....</b>	<b>111</b>
5.4.1 Principal Findings.....	111
5.4.2 Conclusions.....	115
<b>5.5 REFERENCES .....</b>	<b>116</b>
<b>6 MAPPING, INFRASTRUCTURE, AND DATA ANALYSIS FOR THE BRAZILIAN NETWORK OF RARE DISEASES: PROTOCOL FOR THE RARASNET OBSERVATIONAL COHORT STUDY .....</b>	<b>122</b>
<b>6.1 INTRODUCTION .....</b>	<b>122</b>
6.1.1 Background.....	122
6.1.2 Related Work.....	125
6.1.3 Objectives.....	127
<b>6.2 METHODS .....</b>	<b>128</b>
6.2.1 Brazilian Rare Disease Network.....	128
6.2.2 Ethical Considerations.....	130
6.2.3 RARASnet Project Management.....	131
6.2.4 Data Collection Procedures .....	131
6.2.5 Computational Infrastructure and Data Collection Resources .....	133
6.2.6 Database Modeling .....	134
6.2.7 Data Quality Assurance .....	135
6.2.8 Data Management .....	136
6.2.9 Portal Development and Data Analysis.....	137
<b>6.3 RESULTS .....</b>	<b>138</b>
<b>6.4 DISCUSSION .....</b>	<b>140</b>
6.4.1 Main Problems Anticipated and Proposed Solutions .....	140
6.4.2 Applicability of the Results.....	141
6.4.3 Plans for Validation, Dissemination, and Use of Project Results .....	142
<b>6.5 REFERENCES .....</b>	<b>143</b>

<b>7 A PROPOSAL FOR A SET OF ATTRIBUTES RELEVANT FOR WEB PORTAL DATA QUALITY: THE BRAZILIAN RARE DISEASE NETWORK CASE.....</b>	<b>152</b>
<b>7.1 INTRODUCTION .....</b>	<b>152</b>
<b>7.2 RELATED WORK .....</b>	<b>153</b>
<b>7.3 METHODS .....</b>	<b>154</b>
7.3.1 <i>Study design and participants .....</i>	<i>154</i>
7.3.2 <i>Evaluation Process, Metrics/Measures and Guidelines.....</i>	<i>155</i>
7.3.3 <i>Instruments and tools.....</i>	<i>157</i>
7.3.4 <i>Data Analysis Pipeline.....</i>	<i>159</i>
7.3.5 <i>General Evaluation Framework: RE-AIM.....</i>	<i>160</i>
<b>7.4 RELATED WORK.....</b>	<b>161</b>
<b>7.5 FINAL CONSIDERATIONS.....</b>	<b>162</b>
<b>7.6 REFERENCES .....</b>	<b>162</b>
<b>8.1 INTRODUCTION .....</b>	<b>165</b>
<b>8.2 METHODS .....</b>	<b>166</b>
<b>8.3 RESULTS AND DISCUSSION .....</b>	<b>166</b>
<b>8.4 CONCLUSION.....</b>	<b>168</b>
<b>8.5 REFERENCES.....</b>	<b>168</b>
<b>9 A COMPUTATIONAL INFRASTRUCTURE FOR ANALYZING TUBERCULOSIS RESEARCH DATA IN BRAZIL.....</b>	<b>169</b>
<b>9.1 INTRODUCTION .....</b>	<b>169</b>
<b>9.2 METHODS .....</b>	<b>169</b>
9.2.1 <i>Brazilian Tuberculosis Research Network Ecosystem.....</i>	<i>169</i>
9.2.2 <i>Data Gathering, Infrastructure, Curation, and Analysis Pipeline.....</i>	<i>170</i>
9.2.3 <i>Auxiliary Tools Gathering, Infrastructure, Curation, and Analysis Pipeline.....</i>	<i>171</i>
9.2.4 <i>TBWeb Application for TB Analysis.....</i>	<i>171</i>
<b>9.3 EXPECTED OUTCOMES.....</b>	<b>171</b>
<b>9.4 CONCLUSIONS .....</b>	<b>172</b>
<b>9.5 REFERENCES.....</b>	<b>172</b>
<b>10 PROPOSAL FOR A HEALTH INFORMATION MANAGEMENT MODEL BASED ON LEAN THINKING.....</b>	<b>173</b>
<b>10.1 INTRODUCTION .....</b>	<b>173</b>
<b>10.2 BACKGROUND .....</b>	<b>174</b>
10.2.1 <i>Related Work.....</i>	<i>174</i>
10.2.2 <i>Goals.....</i>	<i>175</i>
<b>10.3 RESEARCH PROTOCOL.....</b>	<b>176</b>
10.3.1 <i>Study design .....</i>	<i>176</i>
10.3.2 <i>Data source .....</i>	<i>177</i>
10.3.3 <i>Literature Review.....</i>	<i>177</i>
10.3.2 <i>Lean Tools .....</i>	<i>178</i>
<b>10.4 DISCUSSION.....</b>	<b>179</b>
<b>10.5 CONCLUSIONS AND FUTURE WORK .....</b>	<b>181</b>
<b>10.6 REFERENCES.....</b>	<b>181</b>
<b>11 AN INFORMATION SYSTEM FOR MONITORING TUBERCULOSIS CASES: IMPLEMENTATION RESEARCH PROTOCOL USING RE-AIM FOR A HEALTH REGION IN BRAZIL.....</b>	<b>184</b>
<b>11.1 INTRODUCTION .....</b>	<b>184</b>

<b>11.2 OBJECTIVES</b> .....	185
<b>11.3 METHODS</b> .....	186
11.3.1 <i>Intervention</i> .....	186
11.3.2 <i>Study context</i> .....	186
11.3.3 <i>Pre-implementation planning</i> .....	187
11.3.4 <i>Implementation</i> .....	187
11.3.5 <i>Implementation analysis using RE-AIM</i> .....	188
<b>11.4 FINAL CONSIDERATIONS</b> .....	190
<b>11.5 REFERENCES</b> .....	191
<b>12.1 SHARING AND IMPLEMENTATION RESOURCES</b> .....	195
<b>12.2 COMMUNICATIONS AND DISSEMINATION RESOURCES</b> .....	197
<b>13 CONCLUSIONS AND FUTURE RESEARCH</b> .....	209
<b>13.1 CONCLUSIONS</b> .....	209
<b>13.2 LIMITATIONS</b> .....	214
<b>13.3 FUTURE RESEARCH</b> .....	215
<b>REFERENCES</b> .....	217
<b>APPENDIX A – BRAZILIAN RARE DISEASE PORTAL COMPUTER PROGRAM REGISTRATION CERTIFICATES</b> .....	227
<b>APPENDIX B – SISTB COMPUTER PROGRAM REGISTRATION</b> .....	228



## 1 INTRODUCTION

The rapid evolution of digital technologies in healthcare has revolutionized how we conduct research and approach public health challenges. The growing adoption of digital solutions, such as electronic health records, mobile applications, and online data collection platforms, offers a vast range of opportunities for health research (MOURA JÚNIOR, 2021). However, this digital transformation also underscores the urgent need to establish effective governance models in digital health research. The complexity and scope of data generated by these technologies require integrated approaches that ensure quality, safety, ethics, and effectiveness in a study (CANADA HEALTH INFOWAY, 2016).

In this context, the present research proposes an innovative model of data governance in digital health, aiming not only to consolidate existing practices but also to advance in creating a framework that harmonizes the diversity of initiatives and promotes the ethical and practical use of digital technologies for health research advancement (WORLD HEALTH ORGANIZATION - WHO, 2015). This initiative arises from the critical need to establish clear and comprehensive guidelines to guide the conduct of digital health research, considering not just data collection and analysis but also the translation of these results into practical and effective actions for public health (WANG; BLOCH, 2023).

Therefore, it is necessary to study different application scenarios and their respective governance models, aiming to consider constraints that replicate real systems and problems. The focus is on developing management strategies that are not only theoretically sound but also practical and feasible. These strategies must be effective in outcomes, adaptable to different situations, easy to implement in various contexts, and sustainable in the long term (GLASGOW *et al.*, 2019).

In this sense, the model definition becomes a linchpin in verifying quality and implementation processes within specific application areas, like the health sciences. This study and results emerge from this need and the doctoral candidate engagement in performing their roles as managers of several research projects involving multiple researchers from Brazilian institutions, including universities and health reference services. All these research projects, as will be detailed, were conducted within the Health Intelligence Laboratory (in Portuguese: Laboratório de Inteligência em Saúde – LIS), coordinated by Prof. Domingos Alves, registered and certified in the CNPq's directory of groups (<http://dgp.cnpq.br/dgp/espelhogrupo/794752>).

This project proposes solutions to the problems of digital governance used in

health research. This issue is not as extensively addressed in the literature as other classic problems in the field (KOSTKOVA, 2015). So this doctoral thesis aims to create and make available a governance model for digital health research. The objective is to present methods for problem resolution like digital data infrastructure planning and implementation, data quality techniques, visualization and dissemination of information, and evaluation and monitoring of technological implementation strategies.

The prelude of this doctoral thesis was significantly influenced by the Bernardi *et al.* (2019) paper, which initially focused on data security in digital health. This work led to a paradigm shift in the research scope, expanding it from solely data security to a broader range of digital health governance dimensions. Recognizing data security as just one aspect of data quality, the thesis aims to develop a comprehensive governance model for digital health research.

The uniqueness of this research lies in its pioneering approach towards digital health governance. Unlike conventional studies, our investigation goes beyond data security concerns and delves into a broader spectrum of digital health governance dimensions. Developing a comprehensive governance model, stemming from synthesizing two fundamental protocols, represents a distinctive and innovative contribution to the field.

To thoroughly cover each research topic, this document begins by examining the motivation behind the study and exploring the reasons and importance of the research in the following subsections. This is followed by a delineation of the main and specific objectives the study aims to achieve. The first chapter concludes with a description of the research methodologies used, outlining the process and approaches implemented.

Chapter 3 establishes the study's theoretical basis, delving into digital health globally and within Brazil. It then explores the Findability, Accessibility, Interoperability, and Reuse (FAIR) Principles of digital assets, focusing on data management and stewardship standards (WILKINSON *et al.*, 2016).

It also examines the RE-AIM framework, critical for assessing and planning public health interventions in terms of reach (R), effectiveness (E), and maintenance (M)—which operate at the individual level (i.e., those who are intended to benefit), and adoption (A), implementation (I), and maintenance (M), which focus on the staff and setting levels (GLASGOW; VOGT; BOLES, 1999; GLASGOW *et al.*, 2019). The chapter concludes by presenting the conceptual model central, qualified to the research.

Chapters 4 to 11, encompassing the articles previously mentioned, form the core



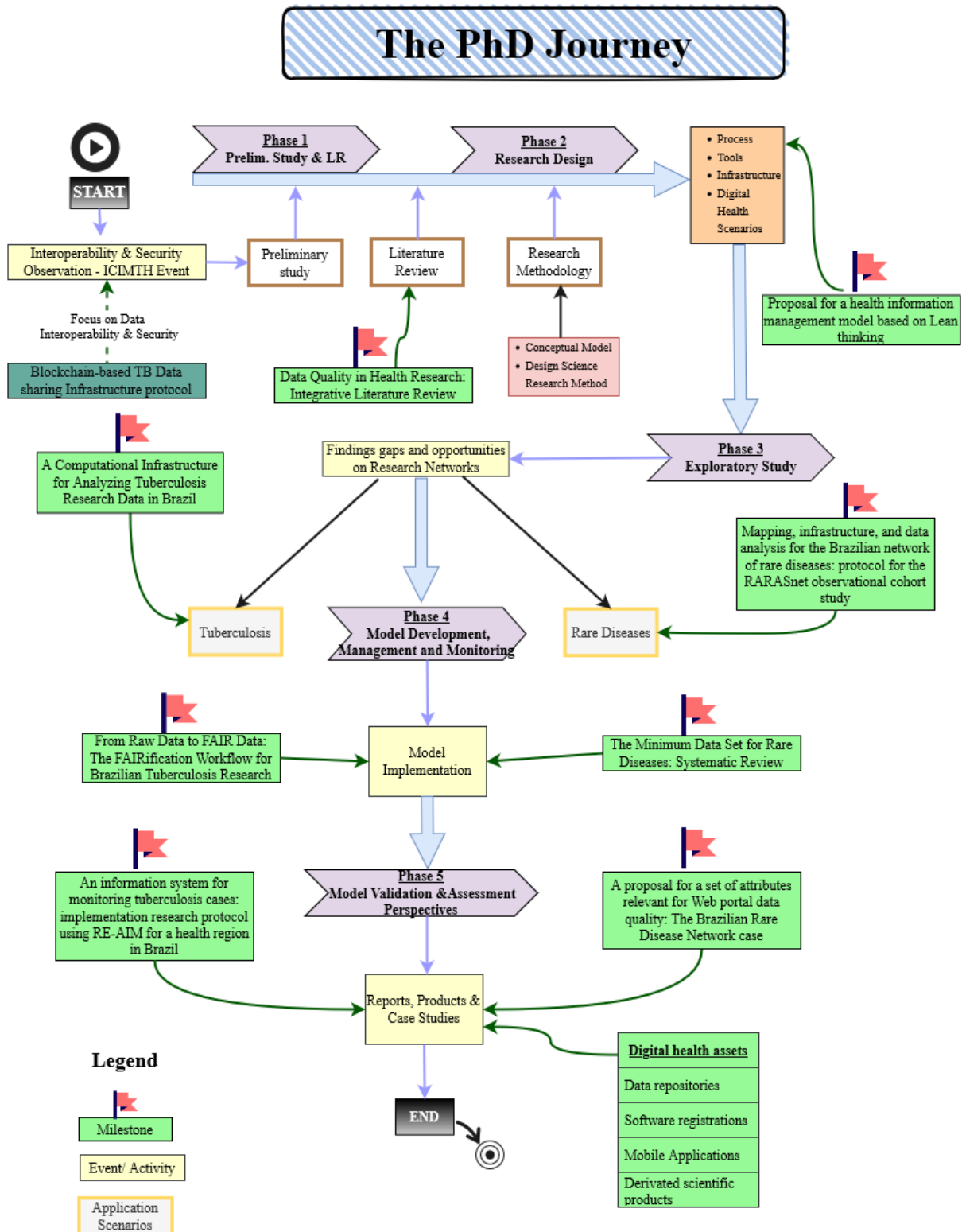
empirical research of the thesis. The author plays a pivotal role in these chapters, demonstrating leadership and managerial responsibility in guiding the research studies. The author's involvement in applying innovative methods and overseeing the research process is prominently featured. These chapters offer comprehensive studies and findings crucial to the research topic.

The author's work meticulously investigates health data quality and management, specifically focusing on rare diseases (RD) and tuberculosis (TB) research in Brazil. The role of the author extends beyond mere oversight; it involves direct engagement in the strategy selection, the application of methodologies, coordination of research activities, and synthesis of findings, highlighting a multi-dimensional exploration of digital health and showcasing the importance of leading significant research initiatives within the field.

Chapter 12 of the dissertation focuses on the open-access dissemination of research content, categorized into distinct groups covering sharing and implementation resources, and communication and dissemination products. This chapter details the extensive range of materials and tools used, from data management plans to source codes, and highlights the distribution of these resources across various data-sharing platforms, demonstrating a commitment to open science. Additionally, it discusses the impact of this thesis-derived scientific output and collaborations, including articles, mentorship activities, and conference papers, on advancing digital health knowledge.

Finally, chapter 13 synthesizes the thesis's key contributions and sets forth proposals for future research, underscoring the ongoing relevance and impact of the study in the digital health field. Presented is the Figure 1 of the Ph.D. Journey, a visual representation encapsulating the comprehensive journey and key milestones of the doctoral research process. The figure identifies critical research phases and the strategies applied across two health domains, adding a layer of practice to our approach.

Figure 1 - PhD Journey



Source: The Author

## 1.1 Motivation

Over a century ago, the history of scientific research was organized with philosophy and achieved notable and excellent development, particularly in the 19th and 20th centuries. Different forms and concerns about explaining the unknown emerged during this period, from understanding man at the mercy of nature and death, from mythic to religious, to philosophical and rational (SILVA, 2001). In light of the transformations the world has been undergoing, strongly driven by the COVID-19 pandemic, challenges related to scientific research are growing, setting new paradigms for researchers in different areas of knowledge (SANTORO, 2020).

In this regard, the literature characterizes terms such as multidisciplinary, interdisciplinary, and transdisciplinarity. In multidisciplinary research, each involved discipline employs its own concepts and methods, sharing only the main research object. Meanwhile, interdisciplinary research examines a particular problem from different perspectives, thus considering different aspects simultaneously. Finally, transdisciplinarity involves a team of researchers with an interdisciplinary spirit without imposing their own ideas (SILVA, 2001).

Therefore, it is known that the use of terms and discourses originates from the reality in which the terminology is employed, reflecting a corresponding historical moment. It's important to note that a term that conveys specific information today may not have had the same connotation over time (KELLER *et al.*, 2017). Given such complexity, combined with constant evolution and paradigm shifts, it is considered that the development of research in an interdisciplinary area like health requires an increasingly close view of the real world (BAMMER, 2013).

An example of this evolution is the increasingly notable popularity of digital health terms among managers, professionals, patients, and researchers. To recognize this topic's importance and reduce its interpretation variability, the WHO launched a guideline on the use of digital interventions in health in 2019. According to these guidelines, the term "digital health" is comprehensive. It includes, in addition to its synonyms, concepts from emerging areas, such as the use of computer and information sciences in big data, genomics, and artificial intelligence" (WHO, 2019a).

With the capacity to substantially affect people's quality of life and even the entire planet's economy, health is, therefore, considered one of the world's most expensive, complex, and critical services. Consequently, it inevitably affects everyone at some point in life, thus

stimulating the perception that the topic is the cornerstone of individual and social well-being (ALKIRE *et al.*, 2020). As such, WHO still reiterates that more people-centered health services are needed (WHO, 2016). This vision directly links with the offerings of emerging digital technologies. However, we still need to fulfill their real potential in the era of people-centered care, and they currently need higher acceptability (ALKIRE *et al.*, 2020; KERNEBECK *et al.*, 2021).

In light of the COVID-19 pandemic, where physical and social distancing of people, including those who are capable of helping them, such as health professionals (LIU *et al.*, 2020), and following such a paradigm shift in the last decade, there has thus arisen, in addition to a growing application of patient-centered management models, a new concept of health system management (SANTORO, 2020). Furthermore, there is an inherent need for digital measures in the face of a global shortage of medical professionals (LIU *et al.*, 2020). Therefore, one of the main pillars of these strategies must contemplate prevention and patient-centered care with the minimization of resource expenditures and the shared use of quality data and information for patient care (WHO, 2016).

The sharing of health data within a person-centered digital health ecosystem and for public interest purposes should be encouraged with patient consent to protect patients' privacy from systems and against inappropriate use. This sharing is vital as it can improve the quality of healthcare processes, health service outcomes, and the continuity of patient care (BRASIL, 2020a).

It can also lead to constructing a knowledge base, which should be capable of interacting with other data systems. The secondary use of health data is vital to improve healthcare quality and research efficacy. Aligned with appropriate digital solutions, it can enable testing, validation, and comparison of artificial intelligence solutions and big data analytics across various parameters and settings, thus continuing patient care (NIH, 2021).

Such interaction requires practices capable of dealing with cybersecurity, trust building, accountability and governance, ethics, equity, empowerment, and literacy, ensuring that good quality data is collected and later shared to support the planning, recognition, and transformation of services, thereby establishing a process of transparency and effective communication about data governance strategies (BRASIL, 2020a).

In addition to governmental and organizational initiatives, various research networks have absorbed these concerns and increasingly developed their research based on plans advocating for solutions based on digital governance models and data quality improvement (JUAREZ *et al.*, 2019). These collaborative networks provide the infrastructure and database

for various types of research, as well as a framework that allows the alignment of a research schedule based on topics that underpin patient and expert concerns about certain circumstances (BIAN *et al.*, 2020).

Although several interventions have shown promise in reducing gaps between evidence and practice, many unanswered questions still need to be answered on rapidly disseminating successful change strategies. With the potential to determine interventions and strategic combinations adjusted to context, research networks can elucidate facilitating aspects or barriers to successful implementation (KHARE *et al.*, 2017). The sharing of quality information and network involvement of many sites and researchers, therefore, offers an opportunity to evaluate and compare the effectiveness of strategies on a large scale and disseminate isonomic and globalized knowledge (WEISKOPF; WENG, 2013).

Difficulties in sharing data may occur for non-technical reasons, such as a lack of trust between data providers. Trust relationships often exist between in-network providers or healthcare organizations. Still, they are particularly difficult to maintain due to the lack of consensus the data can generate if the parties do not use the same healthcare system with a shared provider directory (ZHANG *et al.* 2018).

In this context of widespread diffusion of strategies, data sharing must follow a set of guidelines, architectures, and computational resources (BERNARDI *et al.*, 2019) to address issues relating to governance, standards, and quality of health data from a digital perspective (ZHANG *et al.*, 2018)., as well as to refine the translation and management of scientific research data to contribute to reducing gaps between evidence and practice.

## **1.2 Objectives**

This research aims to create and make available a governance model for digital health research data in order to operationalize the translation of research results into an integrated framework for planning, implementation, and evaluation of various types of public health initiatives and to improve the quality of the collected data and, thereby, the results of the conducted research.

### **1.2.1 Specific objectives**

With the expectation of aligning research outcomes more closely with the real world

through the integration of quality information to enhance interpretation and the sustainable use of data originating from information systems or research projects, this study aims to contribute to reducing the stigma and fear surrounding digital interventions in health. By employing a digital health research governance protocol based jointly on the FAIR principles and the RE-AIM framework, the goal is to provide integrated tools for collecting, managing, and evaluating data quality in health scientific research. The research sets out the following specific objectives to achieve these goals:

- A. Conduct a literature review on data quality in health;
- B. Propose the adoption of frameworks and tools, combined with traditional governance and information exchange techniques, to promote a data quality model for scientific research adopting the concept of digital health in their areas and research networks;
- C. Propose a comprehensive set of data quality mechanisms;
- D. Develop a tool with guidelines for data collection, management, and evaluation in scientific research;
- E. Propose approaches to support the sharing and reuse of data collected in scientific research;
- F. Propose an implementation research model based on guiding principles and guidelines of digital health;
- G. Promote and present the process of adherence, accreditation, and integration of a Brazilian research network in international research consortia engaged with implementation research principles.

### **1.3 Research methodology**

This mixed-methods study comprises research conducted in various mixed-nature scenarios. Mixed-methods research combines qualitative and quantitative research methods to generalize qualitative results, deepen the understanding of quantitative results, and corroborate both qualitative and quantitative outcomes (GALVÃO; PLUYE; RICARTE, 2017). Among the qualitative methods addressed in this thesis are case studies and qualitative descriptions. The quantitative approaches include descriptive cross-sectional research and case study comparisons.

The research aims to be classified as an exploratory and explanatory study, as it has developed and adapted digital strategies to solve proposed problems based on bibliographic studies about problem-solving methods. Thus, this doctoral project's main characteristic is

applying the scientific method through computational techniques as the principal approach to solving fundamental research problems, such as defining and using data collection tools, standards, and guidelines and managing and monitoring data.

## 2 CHAPTERS RELATIONSHIP

Acknowledging the comprehensive exploration of digital health research governance from data quality and implementation research perspective. The thesis systematically navigates the theoretical foundations of digital health, encompassing the FAIR principles and the RE-AIM framework, to the pragmatic applications in health research data management. Thus, Chapter 3 lays the groundwork by discussing digital health in a global and Brazilian context that subsides the presenting conceptual model. This chapter sets the stage for the subsequent empirical investigations, providing a robust theoretical framework that guides the research's direction.

Aiming to critically evaluate and propose methodologies for enhancing data quality in health research from the conceptual model perspective, each chapter builds upon the next, showcasing the development of tools, frameworks, and strategies to optimize digital health research data governance. Chapters 4 to 11 form the core empirical research and demonstrate innovative methods to investigate health data quality and management in application, specifically focusing on RD and TB research in Brazil.

For instance, Chapter 4 begins the empirical journey by conducting an integrative literature review on data quality in health research, identifying key factors affecting data quality and the variability in assessment methods. This review provides a foundational understanding that informs the development of data quality models and frameworks in later chapters. Chapters 5 through 7 progressively build on this foundation by proposing a comprehensive set of data quality mechanisms, developing tools for data collection, management, and evaluation, and proposing approaches to support the sharing and reusing data collected in scientific research. These chapters illustrate the application of the conceptual model and theoretical frameworks established in Chapter 3, showcasing how they are applied to address real-world problems in health research.

Chapter 8 highlights the challenges in TB data management and proposes FAIRification methods, demonstrating the practical application of the FAIR principles. This chapter, along with Chapter 9, which focuses on computational infrastructure for analyzing TB data, illustrates the thesis's commitment to enhancing data management processes and the reusability of data in health research. Chapter 10 introduces a health information management model based on Lean thinking, offering insights into optimizing health data management processes. This novel perspective highlights the thesis's innovative approach to addressing the challenges of digital health research.



Chapter 11 ties together the various aspects of health data management and quality discussed in previous chapters by detailing an information system for monitoring TB cases using the RE-AIM framework in a practical setting. This demonstrates the real-world applicability of the earlier theoretical concepts and underscores the importance of a systematic approach to health data management and research. Finally, Chapter 12 focuses on the open-access dissemination of research content, categorizing into distinct groups covering sharing and implementation resources and communication and dissemination products. This chapter reflects the thesis's commitment to open science and the dissemination of research findings, tools, and methodologies developed throughout the thesis.

Through this detailed exploration, it becomes evident that the thesis interweaves theoretical principles with empirical research across published papers presented in chapters 4 to 11. The correlation lies in the continuous application of the FAIR principles and RE-AIM framework to address specific challenges in health data quality and management, culminating in practical solutions that enhance the governance of digital health research. This integrated approach not only advances the field of digital health research but also provides a scalable model for addressing similar challenges in other research domains.

## 3 THEORETICAL FOUNDATIONS

The theoretical foundation chapter will describe the principles and framework of data management and stewardship guidelines to be used in the doctoral project, and a summary of these applications state-of-the-art applied to the classical digital health scenarios. In addition, it will describe the conceptual model of application of these combined methods that have not yet been performed for health research problems, constituting a gap to be addressed in research.

### 3.1 Digital Health

#### 3.1.1 Digital Health on the World

Although some authors now view health technology as a new era within the fourth industrial revolution, it is not considered a panacea for all health-related issues (HERSH, 2004). Rowlands (2019) categorizes the emergence of information technology (IT) in healthcare systems into four distinct periods:

1. Technology in health, initially limited to corporate support functions (early 1950s).
2. A focus on performance, logistics, organizational functions, and management software (early 1970s).
3. An effort in patient care and digital technologies centered on provider-directed and controlled care processes (the eHealth model - early 2000s).
4. A person- and patient-centered focus, where care aligns with lifestyle and employs new data sources (from 2020 onwards).

Owing to its natural association with predecessors and synonymous terms such as telemedicine, m-health, and e-health, the current definition of digital health has varied and posed an obstacle to research, policy, and practice in this field. According to Oh *et al.* (2005), there are 51 definitions of e-health, generally used in contexts referring to services and systems rather than people's health. Despite various uses of the term, all are directly or indirectly related to technology. Similarly, a quantitative analysis and term mapping of recently published definitions of digital health demonstrated that, despite 95 unique definitions in academic and general sources, their use is more concerned with healthcare delivery than technology use (VARRI *et al.*, 2020).

Among such definitions, the United States Food and Drug Administration (FDA) considers the term to encompass "a broad scope that includes mHealth, health information technology, wearable devices, telehealth, telemedicine, and personalized medicine,"

utilizing computing platforms, connectivity, software, and sensors for healthcare and related uses (FDA, 2019). Meskó *et al.* (2017) further argue that it represents a cultural transformation of revolutionary technologies capable of enabling accessible digital objectives and data, thus promoting a shared decision-making relationship of parity between professionals and patients.

However, observers also perceive this democratization of care as presenting a worrying risk that could reverse the fragile gains in human rights. These gains have been carefully incorporated, for example, into the global response to HIV and tuberculosis (TB) (WHO, 2013). The sharing and commodification of sensitive health and behavioral information could expose vulnerable communities to heightened risks of reprisals and undermine, for instance, the right to non-discrimination in access to healthcare services, the right to privacy, and the principle of inclusive governance. Data enabling precise location and identifying individuals or groups from such populations by state or non-state actors increase the risk of discrimination, imprisonment, and violence (DAVIS *et al.*, 2020).

On the other hand, collecting metadata from mobile phone messages can allow the same actors to infer details such as sleep patterns, travel routines, or frequent contacts, enabling profiling or targeting of vulnerable groups in conflict environments. The debate over the risk-benefit ratio of applying such digital measures to informed consent, especially for those with limited access to healthcare services, has been a global concern (DAVIS *et al.*, 2020).

Questions like "Can global health agencies' data protection policies be maintained if governmental actors demand access to volatile data about stigmatized or criminalized groups, especially due to weak data protection laws and weaker enforcement in many countries?" have been widely highlighted. Despite this, we need to clarify how to address such concerns and, most importantly, whether the affected communities are participating in developing the digital strategy or being consulted in deploying new tools (DAVIS *et al.*, 2020).

In 2005, the World Health Assembly advised all Member States through a resolution to "consider the development of a long-term strategic plan to develop and implement eHealth services to develop the information and communication technology infrastructure for health in order to promote universal access to its benefits" (WHO, 2012).

As a result, more than 120 countries, including low- and middle-income countries, have developed strategies and policies aiming to direct their efforts towards creating a

consistent vision of eHealth aligned with a country's health priorities and resources, developing an action plan to fulfill the proposed vision, and creating a framework to monitor and evaluate the implementation and its progress (WHO, 2018a).

In 2013, the addition of recommendations on standardization and interoperability aimed to cover the development of policies and legislative mechanisms linked to a general national eHealth strategy. In this context, the World Health Organization (WHO) coined "digital health" as an area strongly linked to eHealth, conditioning it to IT and communication in support of health and its related areas (WHO, 2012).

In March 2019, following discussions in online public forums, technical consultations, and WHO regional committee meetings, a consultative process to strengthen digital health implementation announced guidelines and priority recommendations on digital interventions for strengthening the health system. In the same year, the WHO established a digital health department and launched a guideline on digital health intervention (WHO, 2019a).

Also, in 2019, WHO released the first evidence-based guidelines to provide recommendations on selected digital health interventions involving the use of mobile devices (WHO, 2019b). Among them, information on implementation considerations, quality and certainty of existing evidence, factors related to the acceptability and feasibility of the intervention, and gaps in the evidence that may inform future research, digital tracking of patient health status and service, and digital delivery of training and educational content to healthcare professionals (WHO, 2021).

Considering the fundamental role digital health can play in supporting health systems, particularly in the COVID-19 pandemic, the 73rd World Health Assembly endorsed the global digital health strategy 2020-2025. These guidelines are expected to help provide a roadmap for governments and policymakers in introducing and scaling up digital health interventions to support population health outcomes (LABRIQUE *et al.*, 2020).

Besides potentially increasing access to services and information, new technologies can also expand the work of civil society organizations. Their participation should include plans to empower community monitoring, improve information exchange speed, and lead health agencies to correct the course of strategies in real-time (WHO, 2018a).

Thus, the design of legal environments has subsequent effects on data quality. The data collected by governments are riddled with gaps, especially in marginalized

populations, creating a data paradox where the government's official denial of stigmatized groups contributes to the lack of data on their health needs, reinforcing the absence of services. The underestimation of data in low- and middle-income countries to inform the absolute need for new technologies significantly concerns replicating poor-quality data in algorithm-based decisions, which increasingly exacerbates discrimination (POPPE; SÆBØ; BRAA, 2021).

Groups of experts developing such global guidelines must incorporate community representatives' participation to ensure addressing potential human rights challenges and maximizing potential benefits. They must commit to explaining new technology and being accountable for its use on the ground (LABRIQUE *et al.*, 2020).

One of the main ways technology leads healthcare delivery is through the agility and speed of innovation (LABRIQUE *et al.*, 2020). Its innovations can change how new treatments and interventions are discovered and conducted and, in some cases, implemented more quickly than in an academic research environment, where there are public domain funding constraints (WATTS, 2019). Despite institutional collaboration and its increasing production of peer-reviewed research, patient apprehension regarding data sharing is rising (POPPE; SÆBØ; BRAA, 2021).

Concerns about how data are used and stored have repeatedly spread in the media, resulting in companies striving to innovate in healthcare being branded as cowboys, acting recklessly in a rapidly evolving industry. The case raised important questions about regulation in a new digital era, where access to metadata, such as geolocation, internet search data, and other personal information, can result in the reidentification of patient data, even when provided with de-identified electronic health records (WATTS, 2019).

Stronger and more visible partnerships between the private industry and academia will likely lead to changes in patient perception, especially if patients have the opportunity to question and engage with how companies use their data and receive direct responses (POPPE; SÆBØ; BRAA, 2021). By involving established health researchers and the public at the start of the research process, the industry must offer transparency and build trust, which is essential for health decision-makers to implement innovations in a broader healthcare system. Without trust, the risk perceived by patients will continue to overshadow the potential of innovative solutions to transform healthcare (WATTS, 2019).

### 3.1.2 Digital Health in Brazil

The history of Brazilian health informatization began with the Federal Service for Data Processing (In Portuguese: *Serviço Federal de Processamento de Dados - SERPRO*) creation in the early 1960s and the emergence of the Informatics Center at the Ministry of Health (MH) in 1971. In the 1980s, the National Institute of Medical Assistance of Social Security (In Portuguese: *Instituto Nacional de Assistência Médica da Previdência Social - INAMPS*), responsible for managing the Ambulatory ( In Portuguese: *Sistemas de Informação Ambulatorial - SIA*) and Hospital ( In Portuguese: *Sistemas de Informação Hospitalar - SIH*) Information Systems, and the formation of the Brazilian Society of Health Informatics (SBIS) continued the initiatives at a national level, which until then had been carried out within university communities (BRASIL, 2009; SBIS, 2020).

With the advent of the Unified Health System (In Portuguese: *Sistema Único de Saúde - SUS*) implementation in the early 1990s, three influential episodes contributed to the continuity of our history. The first refers to the creation of the Department of Informatics of SUS (DATASUS) in 1991; the second, six years later, the first federal law regulating the protection of citizen data in databases (no. 9.507/1997); and finally, the following year, the emergence of the Primary Care Information System (SIAB) of the MS (BRASIL, 1997; DATASUS, 2019; SANTOS *et al.*, 2009).

In Brazil, the first document from the Federal Council of Medicine (CFM) addressing a term related to digital health was in 2002, with Resolution No. 1.643 defining the provision of services through telemedicine, with some restrictions. That same year saw the creation of the Brazilian Council of Telemedicine and Telehealth (HARZHEIM *et al.* 2017), followed by the launch of the Telematics and Telemedicine project in 2005 (now the National Telehealth Program Brazil Networks of the MS) and the establishment of the University Network of Telemedicine (RUTE) in 2006 (SILVA, 2012).

A year later, technical norms for digitization and the use of computerized systems for storing and handling electronic health records (EHR) were approved, authorizing the elimination of paper and the sharing of sensitive data (HARZHEIM *et al.* 2017). In 2018, Resolution No. 2.227 authorized teleconsultation and telediagnosis. Then, in 2020, in response to the advancing coronavirus pandemic, the telemedicine law (No. 13.989/2020) came about to support tackling the COVID-19 pandemic (SANTOS *et al.*, 2020).

Various initiatives have expanded beyond distance health practices to regulate and guide the development and adoption of digital solutions. Notable initiatives in the

evolution of Brazilian digital health include the regulations of the National Health Card system in 2011 (BRASIL, 2015), the establishment of rules for the implementation of new applications and systems in SUS in 2013 (BRASIL, 2013), and the launch of the e-SUS Primary Care Strategy, with the institution of the Health Information System for Primary Care (SISAB), replacing SIAB (SOARES, 2016).

MH published ordinance No. 2.073 in August 2011 to regulate health information standards and interoperability in response to the increasing emergence of health information systems. The main objectives of the ordinance include (i) the use of ontologies, terminologies, and codifications for the representation of concepts in health to facilitate the sharing of health information; (ii) enabling functional, syntactic, and semantic interoperability among health information systems; (iii) unique identification of the user across different health information systems (BRASIL, 2011).

The ordinance specifies the use of OpenEHR for defining Electronic Health Records (EHR); HL7 as a standard for interoperability focusing on the exchange of results and test requests; SNOMED-CT for coding clinical terms and mapping terminologies; TISS for health insurance information exchange; HL7 CDA for defining the architecture of clinical documents; DICOM for the exchange of imaging tests; LOINC for coding laboratory tests; ISBT 128 for identifying biomaterial labels; IHE-PIX (Patient Identifier Cross-Referencing) for cross-referencing patients from different information systems; and ISO 13606-2 for interoperability of knowledge models, including archetypes, templates, and management methodology (MACIEL; FERREIRA; MARIN, 2018).

Regarding interoperability, MS, through DATASUS, created a document describing a methodology developed to assist teams in creating software services for MS. The instrument links the phases of a project or service (Initiation, Planning, Execution, Monitoring and Control, and Closure), the disciplines involved in service development (Business Modeling, Analysis, Design, Implementation, Testing, and Deployment), and governance management (Business Process Management - BPM, Data, Infrastructure, IT Corporate Architecture, Projects, Services, Configuration, and Changes) (BRASIL, 2011).

With the model, it is possible to guide the development of services for Service-Oriented Architecture (SOA), Business Activity Monitoring (BAM) solutions, Geoprocessing (GEO) solutions, Business Process Management automation solutions, or traditional software development. An essential fact about the initiatives coordinated by DATASUS is the development methodologies available on its website

(<http://datasus.saude.gov.br/metodologias>), among which the Data Administration Methodology (MAD) and the Mobile Development Management Process (PGDM) can be highlighted (BRASIL, 2020b).

In the context of increasingly detailed guidance on interconnected services and systems, 2014 marked the arrival of the Civil Rights Framework for the Internet (Marco Civil da Internet, MCI), established by Federal Law No. 12.96 as an initial regulatory act to protect personal data. Establishing principles, guarantees, rights, and duties for internet use, the MCI sets forth, within the scope of healthcare, rights for patients in their online interactions with the entire healthcare service provider chain. When not adhered to, these rights lead to severe penalties, such as a maximum fine of 10% of the offending economic group's gross revenue (BRASIL, 2018).

With the absence of a regulatory body and issues of consent violation, the MCI was amended four years later by the General Data Protection Law (in Portuguese: Lei Geral de Proteção aos Dados Pessoais, LGPD) (No. 13.709/2018) (BRASIL, 2020c). Therefore, the confidentiality of records that can identify subjects must be guaranteed, respecting privacy rules under applicable regulatory requirement(s) (ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE - OPAS, 2005). To this end, developing actions to adhere to principles such as transparency and data quality is necessary.

Article 12 of the LGPD states that anonymized data are not considered personal data under the law, except when the process of anonymization and de-identification is reversed (BRASIL, 2020c). The enforcement of the LGPD began a year after its approval with the creation of the National Data Protection Authority (ANPD). However, body consolidation still needs to be implemented to implement the LGPD effectively (MONTINI, 2021).

Regarding financial investments in health technology, the Information Technology Master Plan of the Ministry of Health (in Portuguese: Plano Diretor de Tecnologia da Informação, PDTI) published by MS defines a set of goals and actions that should direct IT investments and initiatives during its validity period, aiming to improve institutional governance and reduce the possibility of inappropriate allocation of public resources (BRASIL, 2020d).

Also supported by the National Policy on Information and Informatics in Health (Política Nacional de Informação e Informática em Saúde, PNIIS), a formalization of federal principles and guidelines, is the document titled "e-Health Strategy for Brazil" (BRASIL, 2016). The Brazilian strategy established based on the model of the National



eHealth Tools released by the WHO (WHO, 2012), builds on a national vision of digital health supported by four pillars: governance and organizational resources, standards and interoperability, infrastructure and human resources.

The strategy also aims to develop an action plan and establish a monitoring and evaluation plan. The "Digital Health Strategy Action, Monitoring, and Evaluation Plan for Brazil 2019-2023 (PAM&A)" seeks to monitor the identification, prioritization, and coordinated integration of programs, projects, and actions in health, as well as information and communication services and systems, in addition to the creation/improvement of financing mechanisms, infrastructure, governance, technologies, and human resources (BRASIL, 2020e).

As part of the PAM&A implementation strategy, the national program "Conecte SUS" stands out, aiming to promote the informatization and integration of healthcare establishments and facilitate access to information for maintaining the care of individuals between the country's public and private network units. To achieve this objective, the creation of the national health data integration platform, called the National Health Data Network (RNDS), and a program to support the informatization and qualification of data in Primary Health Care, titled "Informatiza APS," are included (BRASIL, 2020f).

Despite being based on international models and guidelines, the Brazilian strategy, encompassing all initiatives, experiences, and efforts at the national level to guide, regulate, and promote digital health, has undergone revision. In this context, the "Digital Health Strategy for Brazil 2020-2028" was published in December 2020 (BRASIL, 2020g) with the aim of harmonizing and aligning the different previous ventures, such as the PNIIS, PAM&A, and the 2017 digital health strategy itself. The new plan is also expected to integrate the "National Health Plan" and other country policies, promoting the expected benefits and expectations of participation from public, private, and civil society entities for the systematization and consolidation of the work carried out in digital health over the last decade. Despite these historical milestones, a systematic analysis of the maturity of digital health in the country still needs to be observed in the literature, which is necessary to identify its current state and prospects for evolution.

### 3.2 FAIR Principles

Aligned with the global shift towards open science and open data, the FAIR Principles - Findable, Accessible, Interoperable, and Reusable - can compile considerations to make data minimally locatable, accessible, interoperable, and reusable for humans and computers. It has been promoting data-centered criteria that facilitate increased data sharing among entities, disregarding relationships, power differentials, and historical conditions associated with data collection (FAIRSHARING, 2009).

In the two years following the publication of the FAIR Data Principles in 2016, various efforts across the scientific ecosystem have developed assessment methods for their disciplines using different interpretations and measurement criteria (WILKINSON *et al.*, 2016). This diversity is a benefit for communities attempting to implement FAIR criteria but also a challenge due to the inability to easily compare results (BASAJJA; NAMBOBI; WOLSTENCROFT, 2022).

While other efforts can complement the centralization of data in the FAIR Principles to assist in organizing the responsibilities of data producers and repositories (SANSONE; MCQUILTON; ROCCA-SERRA, 2019), it remains unclear which concepts, techniques, and tools are best suited to honor these principles and translate their outcomes into real-world practices.

With a consolidated knowledge about adopting FAIR principles in the scientific ecosystem, it becomes possible to perceive the different approaches to already studied problems and the quality of the data produced. For research employing more complex digital strategies, such as the adoption of artificial intelligence, there is an opportunity for improvement in various approaches already presented in the literature and even the development of new approaches for new variants of the problem in light of the characteristics and needs present in the real world (SWERTZ *et al.*, 2022).

These principles guide good data and metadata management (LAMPRECHT *et al.*, 2020). Thus, the processes and metrics of design, monitoring, and evaluation involve all the sequential activities to manage a record: (1) Governance; (2) data source; (3) metadata, eCRFs, and standardizations; (4) IT infrastructure; (5) data quality; (6) quality information; (7) documentation; (8) staff training; (9) and data quality auditing (KODRA *et al.*, 2018).

At a conceptual level, the principles encode best practices agreed upon by managers and stakeholders, offering an open and inclusive ecosystem for individuals, institutions, and organizations working together through implementation networks. The implementation

networks act on three pillars of activities: GO CHANGE (Culture), GO TRAIN (Training), and GO BUILD (Technology). In other words, the implementation networks are the main drivers of the GO FAIR initiative through open, inclusive, community-led consortia, self-governed by researchers working across various disciplines and countries (GO FAIR, 2021).

### **3.3 RE-AIM framework**

Implementing a new technology requires a change in the way of working, which leads to the creation of new roles and responsibilities and changes existing ones (WHO, 2011). Quality of service infrastructure and dual documentation (on paper and in the new technology) can result in professional satisfaction with its implementation (TILAHUN; FRITZ, 2015). Planning and management are necessary in this process, so it is essential to identify and involve all those affected at this stage. Thus, even where there is evidence, the challenges related to the implementation of a new technology play a significant role in how the implementation and expansion of a proven approach under research conditions will be carried out in real-world conditions (WHO, 2011).

In this context, implementation research (IR) provides evidence of the best ways to support the adoption and optimization of innovations. IR represents a significant interface between the availability of tools, strategies, and interventions and their use within a health system. For this reason, the relevance of IR lies in its role in introducing interventions in public health programs, where the context of implementation can influence their impact (PETERS; TRAN; ADAM, 2013).

Thus, IR helps to understand how interventions can be adapted and implemented to recognize and comprehend health systems and their implementation bottlenecks. It also enables the identification of implementation options for a specific scenario and the promotion of research adoption (WHO, 2019), with a focus on health interventions and their implementation context, thereby stimulating questions about "how" and "why" (WHO, 2011).

To fully capture the impact of a strategy, it is necessary to evaluate not only the impact on participants but also the impact on the organization offering a service and the community at large. Implementation science frameworks help promote the translation of research into practice and are widely used to plan and evaluate the potential or actual impact on public health and the population (CREPALDI *et al.*, 2023).

RE-AIM, introduced by Glasgow in 1999 and updated in 2019, is a tool used to explore the impact of digital technologies, measuring its effect across five essential dimensions for successful implementation: reach, effectiveness, adoption, implementation, and maintenance

(GLASGOW; VOGT; BOLES, 1999; GLASGOW *et al.*, 2019).

In this framework, reach refers to the number or proportion of individuals affected by the intervention. Effectiveness pertains to the outcomes of the intervention (including negative effects on quality of life and others). It includes the reasons for the intervention's success or lack of positive outcomes. Adoption refers to the proportion and representativeness of provider settings that adopt it and the agents of the intervention.

Implementation refers to the elements of the intervention, such as timely and as-intended delivery, the adaptations made, and its acceptance by end-users. Maintenance actively evaluates the sustainability of an intervention over time across the system, provider, and individual levels. All these domains are necessary for the final impact of the new intervention, and each requires a different evaluation approach.

From the FAIR guidelines and the RE-AIM framework focusing on the quality of digital health research, data, processes, and tools available from both, described in the literature, were combined to different scenarios applied in the context of national research. The results of these applications were used to guide the construction of a model capable of mapping and integrating the best practices of the guides, as mentioned earlier (PETERSEN; VAKKALANKA; KUZNIARZ, 2015). The solutions are applicable, especially in contexts with potentially low financial, human, and technological resources (HARRISON; RAHIMI; DANOVARO-HOLLIDAY, 2020).

### **3.4 Conceptual Model**

#### **3.4.1 Conception**

The articulation of the RE-AIM dimensions provides an opportunity to find synergies between the FAIR Principles and their framework, with actions and responsibilities throughout the entire data lifecycle and ecosystems. The incorporation of RE-AIM in data management requires the use of appropriate governance models. For this study, we defined a conceptual model to link the central data quality activities described by Kodra *et al.* (2018) and the FAIR principles to the stages of RE-AIM. Various techniques, processes, and tools in the literature are involved in acquiring, monitoring, and analyzing data quality.

This study's intervention provides information linked to data activities and valuable tools and services for research using digital health concepts. To assemble our conceptual model to meet the FAIR principles and enable an integrated evaluation with the RE-AIM framework, we used the concepts of collaborative and collective problem-solving from Design Thinking,

the principles of operational excellence for performance improvement from Lean Six Sigma, and the agile development concepts from Scrum for the projection and construction of the proposed solution.

In summary, we used Design Thinking to explore and solve problems, applied Lean Six Sigma to enhance operational efficiency and guide our approach to achieving the right results, and employed the agile Scrum methodology to adapt to changes in software conditions (SCHNEIDER, 2017). Recognized as an excellent approach to improving digital health literacy (MESKÓ *et al.*, 2017), we adhered to the four precepts of Design Thinking, encompassing the stages of:

1. Immersion: researchers and experts from the projects mentioned above were presented with literature data referring to the research problem;
2. Ideation: insights gathered in the previous stage were collected, and from a statistical analysis, a synthesis of information was performed to filter the main ideas and thereby refine the core of the proposed model's construction;
3. Prototyping: the solution was tested with the experts, and a technical-operational feasibility study was conducted. The participation of these professionals contributed to the construction and validation of the content of the proposed model;
4. Experimentation: After the conclusion of the final prototype, the model was evaluated using the RE-AIM framework. This step was crucial to ensure that the model not only met the immediate needs of each application scenario but also allowed sustainability and efficiency of the solution throughout all processes and phases of the research.

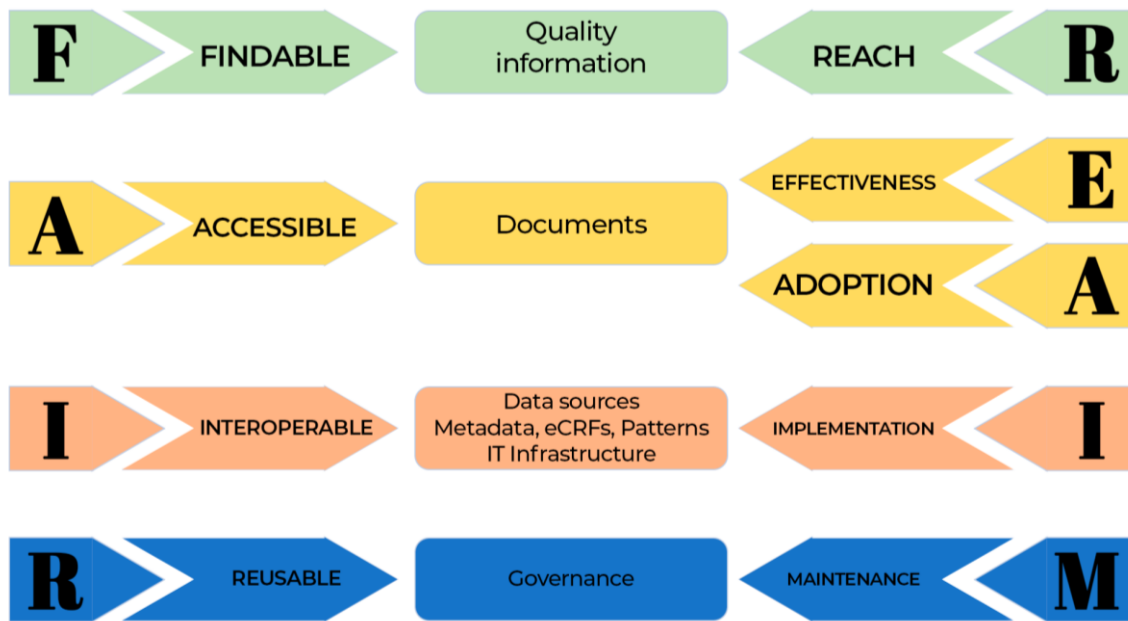
The data and metadata management plan documented all learnings, challenges, and successes during each phase. This documentation was essential for informing future iterations of the model and sharing knowledge with a broader community, contributing to advancements in the field of digital health.

To simplify and harmonize the processes and metrics of design, monitoring, and evaluation that involve all sequential activities for managing a health research registry, it is therefore possible to accelerate the implementation of digital health tools while preparing managers and users for their adoption. In conjunction, we detail a preliminary set of recommendations for the data community to operationalize their digital health research based on the FAIR principles and their activities, combined with the application of the RE-AIM framework and its stages.

The proposed conceptual model reflects and assimilates a set of 9 essential FAIR activities (KODRA *et al.*, 2018), which in turn involve a range of techniques, processes, and

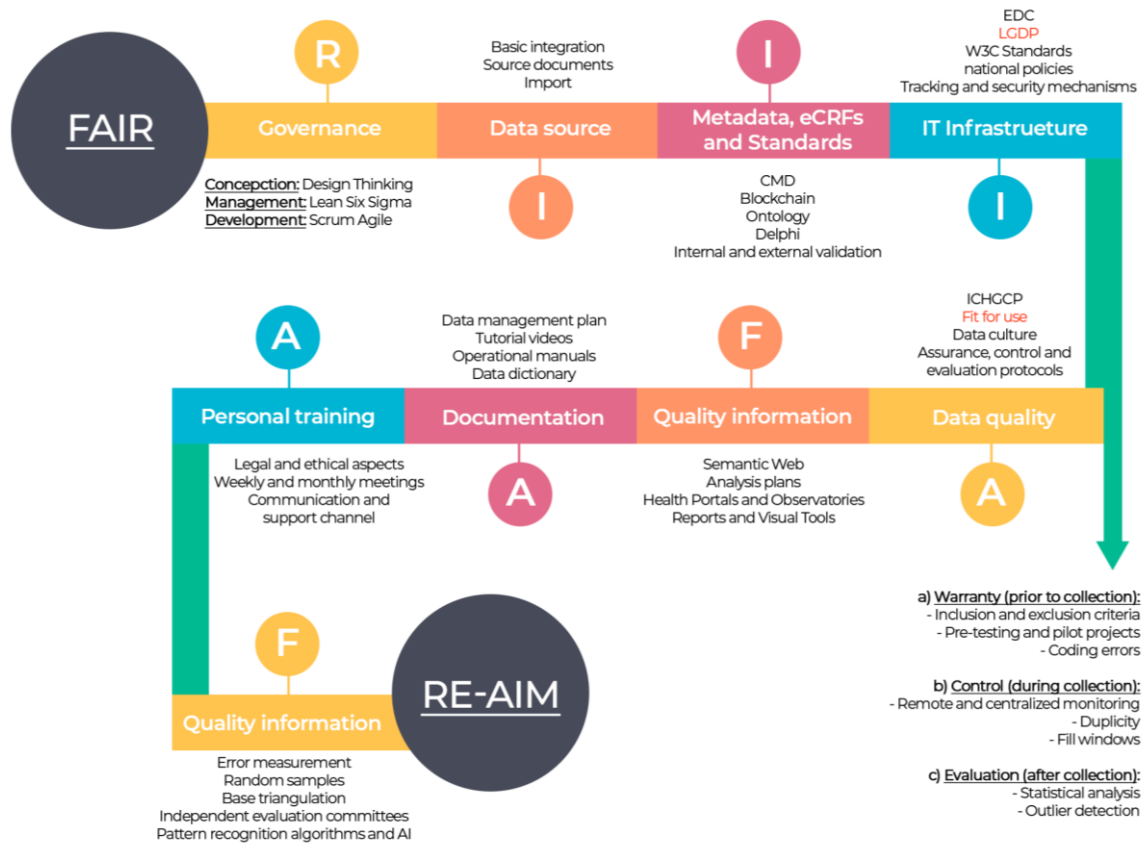
tools present in the literature and also used in the process of data acquisition, monitoring, and quality analysis. Each set of activities in our model reflects each of the five stages, represented by dimensions of the RE-AIM framework. The relationships between FAIR principles and RE-AIM stages are represented in Figure 2 and the activities in the Figure 3.

Figure 2 - The relationships between FAIR activities and RE-AIM.



Source: The Author

Figure 3 - The Activities between FAIR activities and RE-AIM stages.



Source: The Author

### 3.4.2 Application Scenarios

This study directly combines digital health strategies applied in 6 research projects, primarily focusing on Rare Diseases (RD) and TB. Specifically, these projects integrate elements of planning, validation, development, implementation, and evaluation of the practical applicability of digital health solutions for the cross-sectional assessment of governance aspects in scientific research, detailed in the following areas:

#### RD

RD are defined as disorders affecting a small number of people compared to the general population, with a prevalence of less than 65 per 100,000 individuals. In Brazil, reference services for RD were established by the National Policy for Comprehensive Care for People with RD, which offers preventive, diagnostic, and therapeutic actions aimed at reducing morbidity and mortality and improving the quality of life of these individuals (GIUGLIANI *et al.*, 2016). There are over 6,000 RDs that affect more than 8% of the population, therefore

having a significant impact on healthcare systems. The information necessary for evidence-based decision-making must be available to health professionals and managers (BRASIL, 2014).

This information also allows health managers to improve services (BODI *et al.*, 2021). Therefore, the RD field needs increasingly contextualized, precise, and standardized information to reduce the high volume of data fragmentation, creating a solid foundation of information relevant to diagnoses, treatments, and processes (BELLGARD; SNELLING; MCGREE, 2019).

As a result, many countries have initiated national plans to promote care, research, and technology in RD. In the SUS, we observe the existence of a long therapeutic itinerary for RD diagnosis. It can involve patient referrals to different health units of various specialties until the final result, in addition to the difficulty of accessing specialists, examinations, and specific therapies.

Thus, within legal and ethical limits, it is increasingly necessary to use data as comprehensively and efficiently as possible to benefit patients. Important research in the RD field has actively addressed not only the quality of the content of a record but also the quality and mechanisms of how data are used and made available for broader use (RICHTER *et al.*, 2015).

The National Network of Rare Diseases (RARAS) project, approved in the CNPq/MS/SCTIE/DECIT N° 25/2019 notice, aims to conduct the first national representativeness survey on the epidemiology, clinical picture, diagnostic and therapeutic resources employed, and costs in individuals with genetic and non-genetic RDs in Brazil. Among its specific objectives are the consolidation of RARAS with the participation of Reference Services in Rare Diseases (SRDR), University Hospitals (HUs), and Reference Services in Neonatal Screening (SRTNs) of Brazil (to build a national RD database covering all regions of Brazil).

Furthermore, we aim to standardize sociodemographic, epidemiological, clinical (diagnostic and evolution, using international ontology), and therapeutic data throughout the care network. This effort aims to create a national RD registry, providing parameters for studies on the natural history of the disease, disease burden, real-life effectiveness, cost-effectiveness, and budget impact. It will also help identify the ways to access diagnosis and treatment. All this information will be made accessible through a Brazilian Online Atlas of RD, forming the foundation for an observatory and surveillance system for RD.



## TB

*Mycobacterium tuberculosis* (MTB) causes TB, a bacillus that has evolved alongside humans over the years and now stands as the leading cause of death from infectious diseases worldwide. Every year, TB infects approximately 10 million people, leading to 1.2 million deaths (WHO, 2020). As a notifiable disease, health professionals must report every TB case to governmental entities by filling out specific forms or using computerized systems (BRASIL, 2019).

The WHO has established a global strategy for the prevention, treatment, and control of tuberculosis, known as the End TB Strategy, which aims to halt the TB epidemic by 2035 (WHO, 2015). One of the three pillars of this strategy refers to the intensification of scientific research and technological innovation, strongly supporting the reduction of the global TB burden by establishing new diagnostic algorithms, medications, and vaccines and innovative ways of distributing and accessing new resources. Countries with a high incidence of TB can, for example, conduct investigations through their national TB research and control plans and collaborations with national and international partners (WHO, 2018b). Thus, in the realm of digital health applied to TB, the application scenarios involve the following projects:

- **Digital Health for the End TB Strategy: From Integrating Connected Data to Better Evidence-Based Decision-Making:**

Part of the eScience and Data Science Research Program of the São Paulo Research Foundation (FAPESP), the project aims to support healthcare professionals in the 26 municipalities that make up the 13th Regional Health Department of the State of São Paulo (DRS XIII). It facilitates efficient and easy access and monitoring of tuberculosis data. This will be done in collaboration with the Epidemiological Surveillance Group (GVE-XXIV), respecting legal issues advocated by MH and recommendations from the World Wide Web Consortium, with all tasks supported by Semantic Web techniques (LIMA *et al.*, 2021a).

- **Evaluation and Validation of the Performance of Line Probe Assay (LPA):**

The project focuses on applying the rapid diagnostic method for resistant TB and LPA in reference centers in Brazil. It is a commissioned project within the Program of the Secretariat of Health Surveillance of MH, under process number 22410-2019. Its goal is to analyze the diagnostic accuracy of LPA tests using culture and phenotypic sensitivity to antimicrobials as the gold standard. In this project, 7 (seven) municipal or regional Reference Laboratories in

different regions of the country, located in the states of Amazonas, Bahia, Rio de Janeiro, and São Paulo, participate in the clinical-laboratory validation. From its application perspectives, the project expect to analyze the impact on mortality, the proportion of treatment success, and the early systematic detection of resistant TB, as well as to provide an individualized TB regimen using the latest-generation sequencing test (Targeted next-generation sequencing - TNG-S), aided by a decision support system for TB.

- **Development and Validation of Operational Computational Models for the Diagnostic and Therapeutic Cascade of Drug-Resistant TB, through Clinical and Economic Analysis of Targeted Next-Generation Gene Sequencing in Reference Units in Brazil:**

Synergistically to the LPA project, there is the TNG-S project funded by CNPq / MS / SCTIE / Decit from call no. 12/2018 on the theme of research innovation in Health titled "Development and validation of operational computational models for the diagnostic and therapeutic cascade of drug-resistant TB, through the clinical and economic analysis of targeted next-generation gene sequencing, in health reference units in Brazil," process number 440758 / 2018-1. Its objective is to develop software with functionality for clinical care (with electronic patient reports) and epidemiological surveillance (for health information management), integrating multiple predictive models of TB-DR based on artificial intelligence and thus providing TB screening tools and T-NGS analysis.

- **Prospective Multicentric Clinical Trial to Assess the Diagnostic Accuracy of the Truenat Method in Routine Use Indications:**

A pilot study in India on the Truenat MTB test concluded that it achieved high clinical performance, similar to Xpert MTB Rif. However, we need more evidence on the performance of both Truenat MTB and the Truenat MTB-RIF Dx test before recommending their clinical use. Therefore, the Truenat method is being evaluated in different regions of Brazil to confirm its accuracy and ensure that performance characteristics are consistent in routine use and in the Brazilian population, as well as its cost. The prospective, multicentric evaluation of the Truenat MTB tests aims to include 185 adult participants in Brazil with symptoms of pulmonary TB.

- **Validation of Recombinant Purified Protein Derivative (PPD) in TB Infection Diagnosis:**

The expectation of applying the PPD method is to compare the accuracy and costs of the intradermal PPD tests (Russian and Chinese) with the tests currently used in the Brazilian routine for diagnosing latent TB. Low- and middle-income countries concentrate over 80% of the TB burden. Their health professionals diagnose latent TB using the tuberculin test (PT) with PPD Rt-23 (Danish). Even in high-income countries that incorporate interferon-gamma release assays (IGRAs), the PT remains part of the diagnostic algorithm for Latent TB. This prospective, multicentric clinical study across different regions of the country plans to enroll 438 patients with active pulmonary TB for sensitivity analysis and 582 without known TB exposure for specificity analysis. The SVS-MS TB Program contracted the PPD project, under process number 22411-2019, with a total budget of 1 million reais.

### **3.4.3 Process Component**

For the health information management stage, we adopted a model based on Lean thinking, which acts as an auditable instrument for analyzing, representing, and improving the quality of digital health information (TEIXEIRA *et al.*, 2021). Lean IT can be defined as the team's involvement using Lean principles, with systems and tools, incorporating, aligning, and synchronizing IT management with the business area. This approach can offer quality information with effective information systems, promoting continuous improvement and innovation of processes (CANTANHEDE, 2014).

In turn, Lean in healthcare has four mechanisms of change: a) a shared understanding of processes; b) organization aimed at efficiency and effectiveness; c) early error detection and process reliability; d) collaboration for systematic problem-solving and continuous improvement (TEIXEIRA *et al.*, 2021). In this sense, the definition of organized processes and the adoption of continuous thinking in improving the activities performed are fundamental to overcoming the heterogeneity and intangibility present in the spheres that constitute research and health information systems.

For shared understanding, the value stream mapping methods (VSM), Plan, Do, Check, Act (PDCA), and What? Why? Where? When? Who? How? How much? (5W2H) were combined. We evaluated process stages in this approach to determine whether they add value. We aimed to identify potential inflection points for a more comprehensive flow view rather

than just analyzing isolated activities. Thus, when necessary, possible changes in a particular process were described, as well as the moments of implementing changes, observing results, and measures for solving non-conformities (TEIXEIRA *et al.*, 2021).

We then created an action plan, guiding the developed and implemented activities or actions. We reported these to monitor the execution of actions at each planning stage. To organize and represent this plan, we mapped the processes involved in the research using the BPM Notation (BPMN). As a standard for business process modeling, BPMN uses graphical notations to depict the flow's main elements, aiding in identifying and understanding a process's key activities (PULLONEN *et al.*, 2019).

We used the Poka-Yoke technique for error detection. It is a collaborative tool used to avoid defects and prevent potential human errors in workplace processes by applying barriers against errors for warning or control purposes, including detecting human errors or failures (ZHANG, 2014).

#### **3.4.4 Development Component**

Finally, a sociotechnical approach was adopted to carry out a collaborative process in the search for a systematic solution to problems and continuous improvement. In this approach, potential users actively collaborate with technical developers through various strategies, such as regular group meetings, where potential users are motivated to use and improve the proposed system, integrating their workflows and bringing suggestions based on their experience (YOSHIURA, 2020).

We used techniques from the agile development methodology Scrum (KNIBERG, 2015), a dynamic project management tool prevalent in software development and management, to develop computational mechanisms. In this approach, we base projects on fixed-time iterations known as Sprints. Each Sprint represents a period during which we select a set of activities from a task list generated at the beginning of each iteration. Our Team then executes these activities based on their assessment of the time needed to complete various functionalities and prioritize them accordingly (RAMOS; VILELA JUNIOR, 2017).

We conducted weekly Sprints for this project's progress, defining activities and integrating them into the Project Management tool, Trello (JOHNSON, 2017). We assigned each task a specific type, such as "bug", improvement, or feature, along with a title, description, status (new, in progress, resolved), priority level (low, normal, high, urgent, immediate), and

an estimated completion time. We incorporated a tangible portion of the software in each iteration, steadily building towards the final product.

Our Scrum team typically divides roles into three categories: the Product Owner, who creates the system's vision and sets its priorities; the Team, who is responsible for implementing the product; and the Scrum Master, tasked with removing impediments and providing process leadership. The project's support and development team, comprising the advisor, postgraduate students, and collaborators from LIS, fulfilled these roles.

### **3.4.5 Electronic Data Capture Systems (EDCs)**

Although they have limited data-sharing capabilities (METKE-JIMENEZ; HANSEN, 2019), the use of electronic systems for data collection and analysis has been growing over the last decade. Electronic Data Captures (EDCs) eliminate the risks associated with paper instruments and enhance the collection of high-quality data necessary for health research (PASALIC *et al.*, 2018), and can also reduce the cost and time spent on monitoring and data management activities (DILLON *et al.*, 2014).

In this research, we used the REDCap and KoBoToolbox EDCs. We chose these because they are free, stable, widely used, and well-documented software, offering APIs for integration with other systems.

REDCap is a metadata-based web software created in 2004 by a team from Vanderbilt University for conducting classic and longitudinal clinical research, providing researchers with a tool for the design and development of electronic data capture tools (HARRIS *et al.*, 2009; WRIGHT, 2016). REDCap is free software, although not considered open source, which can be installed and managed by a small IT team after obtaining a license (KLIPIN *et al.*, 2014).

Developed by the Harvard Humanitarian Initiative, KoBoToolbox is a free and open-source set of tools for data collection and basic analysis. It was initially built for use in challenging environments in developing countries (HARVARD, 2018). KoBoToolbox uses the Enketo software as a base and offers an environment for building and using forms, online and offline, in any modern browser, thanks to HTML5 features. The software uses the XLSForm standard, simplifying forms' authoring in human-readable spreadsheet format (KHOR *et al.*, 2020). An intuitive form builder serves as an aid tool, allowing researchers to create forms or alternatively create forms in Excel XLS files and import them later.

The REDCap system is widely used by the scientific community to collect and manage

research data, allowing researchers to conduct their studies independently. However, the software can present usability problems during data collection, such as a cluttered graphical interface, gradual performance degradation, and lack of offline operation without relying on a mobile app. On the other hand, although it offers more basic functionalities, KoBoToolbox provides modern styles and allows users to work offline directly from the web browser without needing auxiliary application installation.

### 3.4.6 Semantic Component and Standards

In health research, researchers use semantic annotation to describe the data they collect. Each study typically involves various data collection instruments, encompassing hundreds of fields researchers fill out during the research. Researchers then use annotation to extract and link different sets of research data, which a specific vocabulary describes (SINGHAL; SRIVASTAVA, 2014).

Researchers annotate fields in research forms using semantic vocabularies to enhance the representation of collected data. REDCap allows researchers to include annotations for each field. Although these annotations do not appear on the forms, they remain accessible to the form designer and data exports, improving data interpretability (WRIGHT, 2016). Conversely, KoBoToolbox employs metadata for annotation, mapping the instruments' fields to specific terminologies or ontologies. This approach enables researchers to annotate data as required and utilize it according to their needs.

We employed the REDbox semantic framework to provide a more flexible alternative to direct annotation in EDCs. Developed to improve data collection, management, and sharing in research, REDbox adds meaning to raw data and facilitates the promotion of data quality and availability, especially in low-resource environments (LIMA *et al.*, 2021b).

It is crucial to perform mapping directly in forms or through metadata when defining research data collection instruments. This process should occur alongside selecting sensitive fields, including identifiers and quasi-identifiers. Technically, based on the scope of application, we use available ontologies from BioPortal. BioPortal, as an open database, offers access to biomedical ontologies through web services. This platform facilitates community involvement in evaluating and evolving ontologies and provides additional resources for terminology mapping and establishing review criteria (SALVADORES *et al.*, 2013).

The adopted terminologies include the International Classification of Diseases version

10 (ICD-10), commonly used for diagnoses and procedures (MÖLLER *et al.*, 2010); Orphanet nomenclature (Orphacode), which fits all RDs with a unique code due to its polyhierarchical nature, typically used for more precise diagnoses, clinical findings, and observations (RATH *et al.*, 2012); Online Mendelian Inheritance in Man (OMIM), describing phenotypic characteristics and other information that healthcare professionals involved consider relevant for genetic diseases (HAMOSH *et al.*, 2005).

All three terminologies underwent a mapping and equivalence process by specialists. Typically, terminologies have hierarchical or ontological structures that specify the relationship between terms. Specifically for RDs, as it is an international library, this validation also adapted the definition of certain conditions as rare in the Brazilian scenario since the criterion used for their definition is prevalence.

For cases with absent or suspected diagnoses, we use the Human Phenotype Ontology (HPO) (<https://hpo.jax.org/app/>). By inserting internationally standardized signs and symptoms, healthcare professionals can clinically characterize a participant's condition (ROBINSON; MUNDLOS, 2010). A limitation of adopting HPO in our study is that the HPO list incorporated into REDCap is only available in English. Furthermore, we adopted lists of Brazilian states and municipalities based on the IBGE Municipalities Code Table to accommodate local geographic data.

### **3.4.7 Management and Monitoring Component**

The centralized data management and monitoring process adopted in this study is a continuous remote evaluation of collected and accumulated data, conducted periodically by a team of appropriately qualified personnel designated by the study coordination. Following the international protocol of compiled best practices in human research standards (US DEPARTMENT OF HEALTH AND HUMAN SERVICES, 2021), the defined roles for this activity include the participation of a data manager, coordination, monitors, local focal points where applicable, as well as data entry personnel, often academics and health professionals. The data manager is responsible for maintaining frequent articulation with the coordinating team, local administrators, and collectors, ensuring the quality of data and the security and privacy of research participants.

Researchers participating in this process collaborate with the data manager to establish an operational protocol and determine the source documents (SD) to be used. They conduct remote verification of SD while ensuring the participants' confidentiality is preserved and

without imposing excessive demands on the research center team due to remote monitoring (AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA - ANVISA, 2020a). REDCap, equipped with integrated features, fulfills confidentiality and compliance requirements. The local administrator issues usernames to facilitate access to REDCap.

In this process, researchers collaborate with the data manager to establish an operational protocol and select the source documents (SD) for use. They conduct remote verification of these documents, ensuring participant confidentiality and avoiding excessive workload on the research center team due to remote monitoring (ANVISA, 2020b). REDCap supports this process with its integrated features that adhere to confidentiality and compliance requirements. To facilitate access, the local administrator issues usernames for REDCap.

At the project level, access control is fully configurable: the project owner can set different access levels for the study team, as required by the study protocol (e.g., full access, eligibility control, data entry, monitoring group). Depending on user profiles, identifiers can be flagged and removed from data exports (WRIGHT, 2016).

In the validation and control stage, "cleaning" processes are addressed to generate a high-quality, correct, consistent, and applicable dataset throughout the study. Achieving this involves defining quality and validation rules that actively check for the data's accuracy, completeness, and consistency. These checks can be automated or manual. In REDCap, the Data Resolution Workflow module allows the data manager and monitoring team to implement data quality rules, create queries, and assign them to other project users. Users can respond to queries and mark the data item with a relevant response. After review, the responsible party can close the query (HARRIS *et al.*, 2009).

An open data query indicates an issue with a data value that necessitates resolution. Initiating the query sets in motion the investigation process until the issue is resolved (i.e., closing the query). Any user, regardless of whether the query is specifically assigned to them, can respond to an open query. Data quality issues encompass various problems, including incorrect entries (such as a date significantly deviating from the expected standard), unfilled blank fields, empty mandatory fields, errors in field validation, hidden fields containing data, invalid values in calculated fields (as seen after altering a formula), issues related to temporal consistency, disparities with data collected in other eCRFs, and the identification of extreme values.

The REDCap Data Quality module enables automated checks across the entire database to identify inconsistencies and generate reports for the data manager. Users can open queries from the report and assign them to specific users. Manual checks for inconsistencies can also



be performed for each record and form. The Data Resolution module facilitates the handling of open queries. Access to the module is individually managed for each account or user group (WRIGHT, 2016).

Data modifications can be made by the center's user as long as the record has not been validated and set to read-only status by the manager after validation. Suppose it is necessary to modify data in this state. In that case, the user must request the unlocking of the record via the official project communication channel, which the monitor must-revalidate. Data deletion can only be performed by the manager. Suppose it is necessary to remove the filling of a form, for example. In that case, the user must request the deletion of the record, with the proper justifications, and through the official project communication channel.

Filling and validating records entered into the system involves using three basic markers informed by the data entry user at the end of filling out each form: INCOMPLETE, UNVERIFIED, and COMPLETE. Forms marked as INCOMPLETE refer to a pending action to be taken by the data entry user. Forms marked as UNVERIFIED refer to a pending action to be taken to curate and monitor the data entered by the manager or study monitor.

Forms marked as COMPLETE signify that the associated work is finished without any pending issues, allowing them to be locked for editing. Ensuring the integrity of the data collected in the study necessitates the locking and digital signing (e-signature, when available for the project) of each completed form.

REDCap tracks every operation and allows easy retrieval of this data at the project level, including any modifications to the data collection instruments and data after the project enters production. The teams responsible for data collection received information about all pre-collection phases through training sessions, operational manuals, video recordings, and an official communication channel provided in each project. The remote monitoring procedure, documented and included in the data management plans of the study files, is performed using REDbox documents module (LIMA *et al.*, 2021).

We collect some data that can directly or indirectly identify a participant for specific reasons. We must treat and protect this data primarily. We use the Patient Number to check for record duplication within the center to verify internal duplication and compare it with the source document. For multicentric studies, the Unique Identifier is used to check for record duplication across centers, as given the network articulation of the Unified Health System, it is known that a single participant may be attended in different health units. Thus, identifying duplications across centers ensures that an individual is only considered once during data analysis, thereby avoiding potential biases in data interpretation. Furthermore, in the case of the unavailability of

a unique identifier, the participant's name, along with the mother's name and date of birth, can be used as a composite identifier for detecting duplicity. Both verification strategies can be combined.

## 4 DATA QUALITY IN HEALTH RESEARCH: INTEGRATIVE LITERATURE REVIEW

### *4.1 Introduction*

In health care settings, the priceless value of data must be emphasized, and the relevance and performance of digital media are evidenced by the efforts of governments worldwide to develop infrastructure and technology, aiming to expand their ability to take advantage of generated data. It is important to emphasize that technology, by itself, cannot transform data into information, and the participation of health care professionals is essential for knowledge production from a set of data. Through research that optimizes health interventions and contributes to aligning more effective policies, knowledge combines concrete experiences, values, contexts, and insights, which may enable a framework for evaluation and decision-making [1].

The low quality, nonavailability, and lack of integration (fragmentation) of health data can be highlighted among the main factors that negatively influence research and health decision-making. In addition, it is worth noting the existence of a large number of remote databases accessible only in a particular context. Such factors cause data quality problems and, consequently, information loss. Despite the intense volume, information remains decentralized, but it needs to help the decision-making process [2], making its coordination and evaluation challenging.

The crucial role of data spans a wide range of areas and sectors, ranging from health care data to financial data, social media, transportation, scientific research, and e-commerce. Each data type presents its own challenges and requirements regarding quality, standardization, and privacy. Ensuring the quality and reliability of these data is essential to support the combination of different sources and types of data that can lead to even more powerful discoveries [3].

For example, using poor-quality data in developing artificial intelligence (AI) models can lead to decision-making processes with erroneous conclusions. AI systems, which are increasingly used to aid decision-making, have used labeled big data sets to build their models. Data are often collected and marked by poorly trained algorithms, and research often demonstrates this method's problems. Algorithms can present biases in judgments about a person's profession, nationality, or character and basic errors hidden in the data used to train

and test their models. Consequently, prediction can be masked, making it difficult to distinguish between right and wrong models [4].

Principles are also established in the semantic web domain to ensure adequate data quality for use in linked data environments. Such recommendations are divided into 4 dimensions: quality of data sources, quality of raw data, quality of the semantic conversion, and quality of the linking process. The first principle is related to the availability, accessibility, and reliability of the data source, as well as technical issues, such as performance and verifiability [5]. The second dimension refers to the absence of noise, inconsistencies, and duplicates in the raw data from these data sources. In addition, it also addresses issues regarding the completeness, accuracy, cleanness, and formatting of the data to be helpful and easily converted into other models, if necessary. The last 2 dimensions refer to the use of high-quality validated vocabularies, flexible for semantic conversion, and the ability of these data to be combined with other semantic data, thus generating sophisticated informational intelligence. Such factors depend on correctness, granularity, consistency, connectedness, isomorphism, and directionality [6].

The heterogeneity of data in this area is intrinsically connected to the type of information generated by health services and research, which are considered diverse and complex. The highly heterogeneous and sometimes ambiguous nature of medical language and its constant evolution, the enormous amount of data constantly generated by process automation and the emergence of new technologies, and the need to process and analyze data for decision-making constitute the foundation for the inevitable computerization of health systems and research and to promote the production and management of knowledge [7].

There are different concepts of data quality [8]. According to the WHO, quality data portray what was determined by their official source and must encompass the following characteristics: accuracy and validity, reliability, completeness, readability, timeliness and punctuality, accessibility, meaning or usefulness, confidentiality, and security [9]. Data quality can be affected at different stages, such as the collection process, coding, and nonstandardization of terms. It can be interfered with by technical, organizational, behavioral, and environmental aspects [10].

Even when data exist, some aspects make their use unfeasible by researchers, managers, and health care professionals, such as the noncomputerization of processes, heterogeneity, duplicity, and errors in collecting and processing data in health information systems [11]. Reliable health data must support decision-making and strategies to improve service delivery

to generate consistent evidence on health status, so the data quality management process must ensure the reliability of the data collected [12].

Some health institutions have action protocols that require their departments to adopt quality improvement and resource-saving initiatives. Consequently, various methodologies to improve the quality of services have been applied in the health field. Mulgund et al [13] demonstrated, for example, how data quality from physician-rating sites can empower patients' voices and increase the transparency of health care processes.

Research in scientific communities about new strategies constantly evolves to improve research quality through better reproducibility and empowerment of researchers and provides patient groups with tools for secure data sharing and privacy compliance [14]. Raising a hypothesis and defining a methodology are a standard scientific approach in health research, which will lead to the acquisition of specific data. In contrast, data production in the big data era is often completely independent of the possible use of the data. One of the hallmarks of the big data era is that the data are often used for a purpose other than the one for which they were acquired. In this sense, influencing the modification of acquisition processes in clinical contexts requires more structured approaches [13].

The health sector is increasingly using advanced technologies, such as sophisticated information systems, knowledge-based platforms, machine learning algorithms, semantic web applications, and AI software [15]. These mechanisms use structured data sets to identify patterns, resolve complex problems, assist with managerial and strategic decision-making, and predict future events. However, it is crucial to ensure that the data used for these analyses adhere to the best practices and metrics for evaluating data quality to avoid biases in the conclusions generated by these technologies. Failure to do so can make it challenging to elucidate previously unknown health phenomena and events [16].

To use the best practices, institutions use the results of literature reviews due to the significant time savings and high reliability of their studies. Thus, through an integrative literature review, the main objective of this work is to identify and evaluate digital health technology interventions designed to support the conduct of health research based on data quality.

## ***4.2 Methods***

### **4.2.1 Study Design**

The Population, Concept, and Context (PCC) strategy was applied to define the research question. The PCC strategy guides the question of the study and its elaboration, helping in the process of bibliographic search for evidence. The adequate definition of the research question indicates the information necessary to answer it and avoids the error of unnecessary searches [17].

“Population” refers to the population or problem to be investigated in the study. “Content” refers to all the detailed elements relevant to what would be considered in a formal integrative review, such as interventions and phenomena of interest and outcomes. “Context” is defined according to the objective and the review question. It can be determined by cultural factors, such as geographic location, gender, or ethnicity [18]. For this study, the following were defined: P=digital technology, C=data accuracy, and C=health research.

In this sense, the following research questions were defined:

- What is the definition of health research data quality?
- What are the health research data quality techniques and tools?
- What are the indicators of the data confidence level in health research?

#### **4.2.2 Health Research**

Numerous classifications characterize scientific research, depending on its objective, type of approach, and nature. Regardless of the purpose of how surveys can be classified, levels of confidence in data quality must be ubiquitous at all stages of the survey. Detailed cost-effectiveness analysis may inform decisions to adopt technology methods and tools that support electronic data collection of such interventions as an alternative to traditional methods.

Health research systems have invested heavily in research and development to support sound decisions. In this sense, all types of studies were observed that presented results of recent opportunities to apply the value of digital technology to the quality of the information in the direct or indirect evaluation of the promotion of health research. Therefore, in a transversal way, we considered all types of studies dealing with such aspects.

#### **4.2.3 Types of Approaches**

Various methods for setting priorities in health technology research and development have been proposed, and some have been used to identify priority areas for research. They

include surveys and measurements of epidemiological estimates, clinical research, and cost-effectiveness assessments of devices and drugs. The technical challenges and estimation of losses due to variations in clinical practice and deviations from protocols have been supported by recommendation manuals and good practice guidelines. However, each of these proposed methods has specific severe methodological problems.

First, all these approaches see research simply as a method of changing clinical practice. However, there are many ways to change clinical practice, and conducting research may not be the most effective or cost-effective way. Research's real value is generating information about what clinical practice should be. The question of how to implement survey results is a separate but related issue. Therefore, these methods implicitly assume no uncertainty surrounding the decision that the proposed research should inform.

#### **4.2.4 Types of Interventions and Evaluated Results**

Technology-based interventions that affect and aggregate concepts, designs, methods, processes, and outcomes promote data quality from all health research.

Measures demonstrate how results can address political, ethical, and legal issues, including the need to support and use technological mechanisms that bring added value regardless of the type and stage at which they are applied to research. We looked at how the results can be evaluated to address other questions, such as which subgroups of domains should be prioritized, which comparators and outcomes should be included, and which follow-up duration and moments would be most valuable for improving interventions on the reliability of health research data.

#### **4.2.5 Eligibility Criteria**

Research carried out in English and Portuguese, with quantitative and qualitative approaches, primary studies, systematic reviews, meta-analyses, meta-synthesis, books, and guidelines, published from 2016 onward was included. This choice is justified because we sought scientific indications that were minimally evaluated by our community. In this sense, websites, white papers, reports, abstracts only, letters, and commentaries were not considered. The year limitation is justified because knowledge is considered an adequate degree of being up to date.

In addition to the methodological design, we included any studies that described the definition, techniques, or tools that have the essential functions of synthesis, integration, and verification of existing data from different research sources to guarantee acceptable levels of data quality. In this way, we expected to monitor trends in health research, highlight areas for action on this topic, and, finally, identify gaps in health data arising from quality control applications.

Although the primary objective of this review was to seek evidence of data quality from health research, we also independently included studies on health data quality and research data quality. The exclusion criteria were applied to studies with a lack of information (eg, the paper was not found), studies whose primary focus was not health and research, and papers not relevant to the objective of the research, papers not available as full text in the final search, and papers not written in English or Portuguese. In addition, the titles and respective authors were checked to verify possible database repetitions. All criteria are presented in Table 1

**Table 1 - Inclusion and exclusion criteria for eligibility of studies.**

Category	Inclusion criteria	Exclusion criteria
Approach	Quantitative, qualitative	— <sup>a</sup>
Document type	Primary studies, systematic reviews, meta-analyses, meta-synthesis, books, guidelines	Websites, white papers, reports, abstracts only, letters, commentaries
Year	Starting from 2016	Before 2016
Information	Describe the definition, techniques, or tools that have functions of synthesis, integration, and verification of existing data from different research sources	Lack information, not available as full text
Study focus	—	Not health and research, not relevant to the objective
Language	—	Not in English or Portuguese

<sup>a</sup>Not applicable.

#### 4.2.6 Databases and Search Strategies

A search was carried out in 6 electronic scientific databases in January 2022 because of their quality parameters and broad scope: PubMed, SCOPUS, Web of Science, Institute of Electrical and Electronics Engineers (IEEE) Digital Library, Cumulative Index of Nursing and Allied Health Literature (CINAHL), and Latin American and Caribbean Health Sciences Literature (LILACS). For the search, descriptors and their synonyms were combined according



to the Health Sciences Descriptors (DeCS) [19] and Medical Subject Headings (MeSH) [20]. The following descriptors and keywords were selected, combined with the Boolean connectors AND and OR: “Data Accuracy,” “Data Gathering,” and “Health Research.” These descriptors and keywords come from an iterative and tuning process after an exploratory phase. The same search strategy was used in all databases.

Google Scholar was used for manual searching, searching for other references, and searching for dissertations. These documents are considered gray literature because they are not published in commercial media. However, they may thus reduce publication bias, increase reviews’ comprehensiveness and timeliness, and foster a balanced picture of available evidence [21].

We created a list of all the studies we found and removed duplicates. A manual search was performed for possible studies/reports not found in the databases. The references of each analyzed study were also reviewed for inclusion in the search. The search was carried out in January 2022, and based on the inclusion and exclusion criteria described, the final number of papers included in the proposed integrative review was reached. The search procedure in the databases and data platforms is described in Table 2, according to the combination of descriptors.

**Table 2** - Search procedure on databases.

Database	Search string	Query result (N=27,709), n (%)
PubMed	(“Data Accuracy” OR “Data Gathering”) AND “Health Research”	19,340 (69.80)
SCOPUS	TITLE-ABS-KEY ((data AND accuracy OR data AND gathering) AND health AND research) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016))	789 (2.84)
Web of Science	((“Data Accuracy” OR “Data Gathering”) AND “Health Research”)	5589 (20.17)
IEEEa Digital Library	(“Index Terms”:Data Accuracy) OR (“Index Terms”:Data Gathering) AND (“Index Terms”:Health Research)	1989 (7.18)

LILACSB	Data Accuracy [Palavras] or Data Collection [Palavras] and Health Research Evaluation [Palavras]	2 (0.01)
CINAHL <sup>a</sup>	(“Data Accuracy” OR “Data Gathering”) AND “Health Research”	0

<sup>a</sup>IEEE: Institute of Electrical and Electronics Engineers.

<sup>b</sup>LILACS: Latin American and Caribbean Health Sciences Literature.

<sup>c</sup>CINAHL: Cumulative Index of Nursing and Allied Health Literature.

#### 4.2.7 Data Collection

First, 2 independent reviewers with expertise in information and data science performed a careful reading of the title of each paper. The selected papers were filtered after reading the abstract and selected according to the presence of keywords and descriptors of interest. The reviewers were not blinded to the journal’s title, study authors, or associated institutions. The established inclusion and exclusion criteria adequacy was verified for all screened publications. Any disagreements between the 2 reviewers were resolved by a senior third independent evaluator. The Mendeley reference manager [22] was used to organize the papers. Subsequently, the extracted findings were shared and discussed with the other team members.

Data synthesis aims to gather findings into themes/topics that represent, describe, and explain the phenomena under study. The extracted data were analyzed to identify themes arising from the data and facilitate the integration and development of the theory. Two reviewers performed data analysis and shared it with other team members to ensure the synthesis adequately reflected the original data.

#### 4.2.8 Data Extraction

Data extraction involved first-order (participants’ citations) or second-order (researchers’ interpretation, statements, assumptions, and ideas) concepts in qualitative research. Second-order concepts were extracted to answer the questions of this study [17].

We looked at data quality characteristics in the studies examined, the assessment methods used, and basic descriptive information, including the type of data under study. Before starting this analysis, we looked for preexisting data quality and governance models specific to

health research but needed help finding them. Thus, 2 reviewers were responsible for extracting the following data from each paper:

- Bibliographic information (title, publication date and journal, and authors)
- Study objectives
- Methods (study design, data collection, and analysis)
- Results (researchers' interpretation, statements, assumptions, and ideas)

#### 4.2.9 Result Presentation

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Multimedia Appendix 1) and flowchart were used to visualize the search strategy results in the databases. PRISMA follows a minimum set of items to improve reviews and meta-analyses [23]. Based on the PRISMA flowchart, a narrative synthesis was prepared, in which we described the objectives and purposes of the selected and reviewed papers, the concepts adopted, and the results related to the theme of this review.

#### 4.2.10 Data Synthesis

The data synthesis process involved several steps to ensure a systematic and comprehensive analysis of the findings. After a rigorous study selection process, the extracted data were analyzed using a coding and categorization approach.

Initially, a coding framework was developed based on the research objectives and key themes identified in the literature. This framework served as a guide for organizing and categorizing the extracted data. At least 2 independent reviewers performed this coding process to ensure consistency and minimize bias. Any discrepancies or disagreements were resolved through consensus discussions. Relevant data points from each study were coded and assigned to specific categories or themes (Multimedia Appendix 2), capturing the main aspects related to data quality in health research, as shown in Table 3.

**Table 3** - Search procedure used on databases.

Category	Subtopics/example codes
Data quality assessment methods	<ul style="list-style-type: none"> <li>• Ontologies, adjust to fit, frameworks, guidelines</li> <li>• Quality dimensions</li> </ul>

Factors influencing data quality	<ul style="list-style-type: none"> <li>• Study design, application/data sources</li> <li>• Context, limitations</li> </ul>
Strategies for improving data quality	<ul style="list-style-type: none"> <li>• Process, tools, techniques/analysis</li> </ul>

Once the data were coded and categorized, a thorough analysis was conducted to identify patterns, trends, and commonalities across the studies. Quantitative data, such as frequencies or percentages of reported data quality issues, were analyzed using descriptive statistics. Qualitative data, such as themes or explanations provided by the authors, were analyzed using thematic analysis techniques to identify recurring concepts or narratives related to data quality.

The synthesized findings were then summarized and organized into coherent themes or subtopics. This involved integrating the coded data from different studies to identify overarching patterns and relationships. Similar results were grouped, and relationships between different themes or categories were explored to derive meaningful insights and generate a comprehensive picture of data quality in health research.

As part of the data synthesis process, the quality of the included studies was also assessed. This involved evaluating the studies' methodological rigor, reliability, and validity using established quality assessment tools or frameworks. The quality assessment results were considered when interpreting and discussing the synthesized findings, providing a context for understanding the strength and limitations of the evidence.

### ***4.3 Results***

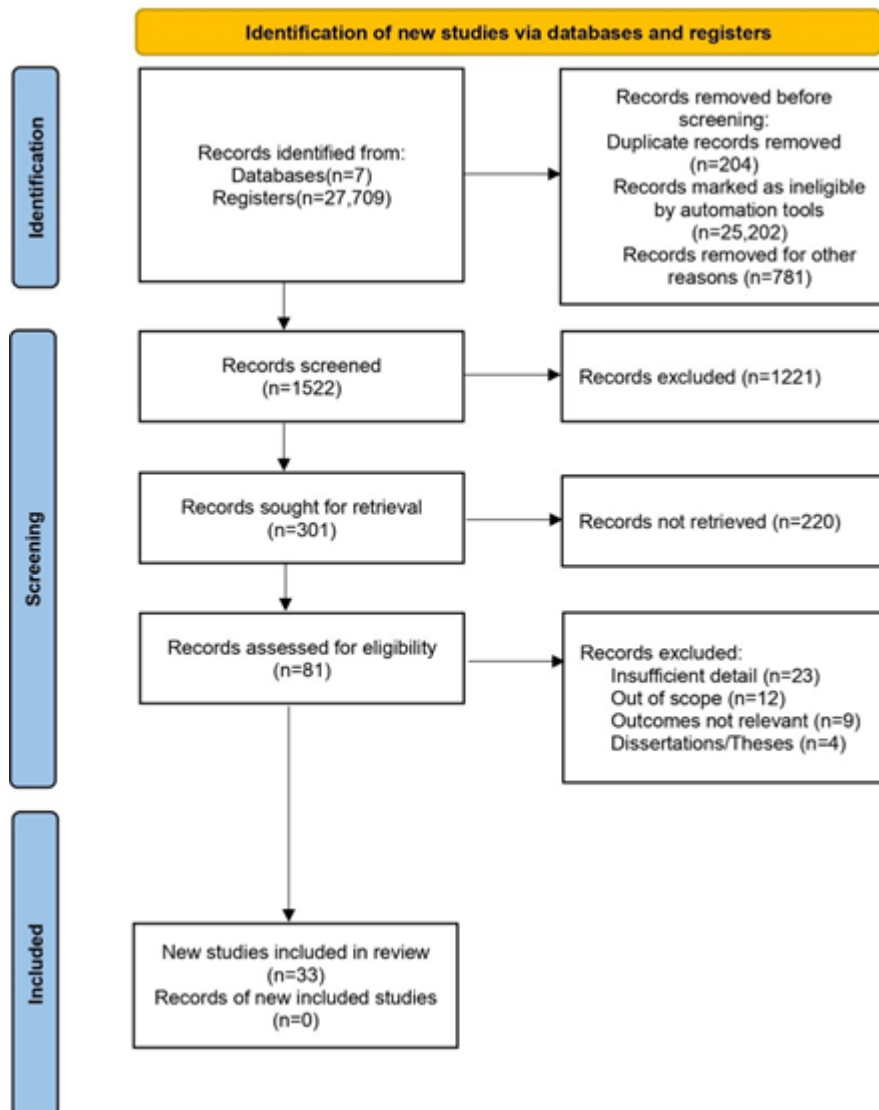
#### **4.3.1 Study Characteristics**

In this review, 27,709 occurrences were returned from the search procedure, with 789 (2.84%) records from the SCOPUS database, 2 (0.01%) from LILACS, 1989 (7.18%) from the IEEE Digital Library, 5589 (20.17%) from the Web of Science, and 19,340 (69.80%) from PubMed. Searches were also performed in the World Health Organization Library and Information Networks for Knowledge (WHOLIS) and CINAHL databases, but no results were found. Of these, 25,202 (90.95%) records were flagged as ineligible by the automation tools and filters available in the databases, because they were mainly reports, editorial papers, letters or comments, book chapters, dissertations, and theses or because they did not specifically

address the topic of interest according to the use of descriptors. Furthermore, 204 (0.74%) records were duplicated between databases and were removed.

After carefully evaluating the titles and abstracts (first screening step), 1221 (80.22%) of 1522 search results were excluded. For inclusion of papers after reading the abstracts, 81 ( ) of 301 (26.9%) papers were listed for a full reading. After analyzing and extracting the desired results, 33 (40.7%) papers were included in the review because they answered the research questions. The entire selection, sorting, extraction, and synthesis process is described through the PRISMA flowchart [23], represented in Figure 4.

Figure 4 - PRISMA flowchart with the results of study selection. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



The 33 studies covered the period of 2017-2021 and were conducted in 22 countries. Most studies were concentrated in Europe (n=11, 33.3%) and North America (the United States and Canada; n=10, 30.3%). Others were carried out in Oceania (Australia; n=4, 12.1%), Asia (China and Taiwan; n=3, 9.1%), and the Middle East (Iran and Saudi Arabia; n=2, 6.1%). In addition, studies were carried out collaboratively or in a network (the United States and India; the United States and African countries; the US Consortium, the United Kingdom, South Africa, Costa Rica, Canada, Sweden, Switzerland, and Bahrain; n=3, 9.1%).

In their entirety, the studies were carried out in high-income countries, and most of the assessments were based on the evidence available in English. The United States (n=11, 33.3%) and Australia (n=4, 12.1%) led in studies involving the investigated topic. No studies conducted or coordinated by middle-income countries were reported. In addition to the low economic diversity of countries where the research was conducted, all papers were evaluated in a single language. The involvement and collaboration of emerging countries took place exclusively through partnerships and participation in consortia.

Regarding the domains described in the studies, there was tremendous variability and inconsistency between the terms presented (n=38 terms). Note that no consensus existed between critical and noncritical variables for data quality assessment. The lack of consensus reflected that the definitions of concepts vary and their relationships are not homogeneous across studies. The discrepancy between domains and evaluated concepts did not allow an evaluation of parity between metrics and was present during all phases of the studies found. The subtopic distribution into the defined categories also evidenced the high-variability factors and strategies in the literature to lead with data quality. The distribution of the improvement strategies for data quality is shown in Table 4 and that of the related influencing factors for data quality in Table 5.

**Table 4 - Improvement strategies for data quality.**

Category and improvement strategy		Distribution of improvement strategy in studies (N=33), n (%)
<b>Process</b>		
	Business intelligence model	12 (36.4)
	Monitoring	7 (21.2)
	Benchmarking	3 (9.1)
	No well-established process followed	11 (33.3)
<b>Techniques/analysis</b>		
	Quantitative	12 (36.4)
	Qualitative	11 (33.3)
	Mixed	10 (30.3)
<b>Tools</b>		
	Minimum data set definition	7 (21.2)
	Audit	13 (39.4)
	Error detection	3 (9.1)
	Decision support	2 (6.1)
	Multiple tools	2 (6.1)
	No tools used	6 (18.2)

**Table 5 - Influencing factors for data quality.**

Strategy and influencing factors		Distribution of influencing factor in studies (N=33), n (%)
<b>Strategy: process; influencing factor category: context</b>		
	Research-only controlled environment	7 (21.2)
	Transition and validation environment	8 (24.2)
	Restricted routine environment	18 (54.6)
<b>Strategy: techniques/analysis; influencing factor category: study design</b>		
	Longitudinal	27 (81.8)
	Cross-sectional	3 (9.1)
	Combined	2 (6.1)
	No information provided/design adopted unclear	1 (3.0)
<b>Strategy: tools; influencing factor category: data source</b>		
	Own research repositories	24 (72.7)
	Preexisting data models	4 (12.1)
	Public databases	2 (6.1)
	Other sources	3 (9.1)
<b>Strategy: tools; influencing factor category: study limitations</b>		
	Methodological	21 (63.6)
	Technical	15 (45.5)
	Social	6 (18.2)
	Organizational	2 (6.1)
	Legal	1 (3.0)
<b>Strategy: tools; influencing factor category: study design</b>		
	Longitudinal	27 (81.8)
	Cross-sectional	3 (9.1)
	Combined	2 (6.1)
	No information provided/design adopted unclear	1 (3.0)



### 4.3.2 Data Quality Issues and Challenges

The metrics extracted from the studies comprised domains related to the methodology adopted by them, that is, concepts that supported the definition of data quality and their respective individual or combined categorizations regarding the adjusted use for the purpose (n=8, 24.2%) of frameworks (n=6, 18.2%), ontologies (n=2, 6.1%), good practice guides (n=15, 45.5%), or combinations of methodologies (n=2, 6.1%).

Among the studies that used the concept of purpose-adjusted use, terms such as “gold standard according to experts” [24], “intrinsic quality” [25], “ideal record” [26], “data fitness” [27,28], and “data culture” [29,30] were addressed. In general, the use of frameworks and ontologies was based on previously published studies and available in development libraries as modules for mapping-adapted entities, proprietary or embedded systems, and data-based strategies for process improvement [31-34].

The central guides and guidelines adopted in data quality studies refer to the adoption of national protocols and policies, agreements signed between research networks and consortia, guides to good clinical practices (International Conference on Harmonization—Good Clinical Practice, ICHGCP [35-38]; Food and Drug Administration, FDA [35,38]; Health Insurance Portability and Accountability Act, HIPPA [39]), or information governance principles, models, and strategies (International Organization for Standardization, ISO [40,41]; Joint Action Cross-Border Patient Registries Initiative, PARENT [41]; Findability, Accessibility, Interoperability, and Reuse, FAIR [25,40,42]).

Regarding data quality, dimensions were interposed in all research stages, thus being a fundamental factor in being incorporated with good practices and recommendations, giving light to health research, regardless of their methodological designs. The distribution of dimensions evaluated in our findings showed significant heterogeneity, as shown in Table 6.

**Table 6** - Distribution of quality dimensions in health research.

<b>Data quality dimension(s)</b>	<b>Distribution in studies (N=33), n (%)</b>
Integrity	19 (57.6)
Precision	16 (48.5)
Consistency	11 (33.3)
Opportunity	10 (30.3)
Validity, plausibility	7 (21.2)
Relevance	5 (15.1)
Accuracy, accessibility, utility, conformity	4 (12.1)
Reliability, trust, interoperability, usability	3 (9.1)
Correctness, comparability, inconsistency, flexibility, security, availability	2 (6.1)
Credibility, incompleteness, bias, variance, frequency, prevention, singularity, temporality, exclusivity, uniqueness, currentness, consent, loss, degradation, simplicity, acceptability, interpretability, coherence	1 (3.0)

### 4.3.3 Factors Affecting Data Quality

Regarding data quality, dimensions were interposed in all research stages, thus being a fundamental factor in being incorporated with good practices and recommendations, giving light to health research, regardless of their methodological designs. The distribution of dimensions evaluated in our findings showed significant heterogeneity, as shown in Table 6.

The study considered factors such as the environment, application time, and development steps, all influencing data quality. Controlled environments were reported in research-only scenarios with planning and proof-of-concept development [34,35,37,38,43-45]. Transition and validation environments were identified where research and service were

combined [25,27,31,40,46-49]. Most studies were conducted in restricted environments specific to health services. Most studies also used their own research repositories, while others relied on external sources, such as preexisting data models [25,26,33,40] or public databases [38,50]. The research applications spanned diverse health areas, including electronic health records, cancer, intensive care units, rare diseases, maternal health, and more. However, the research areas were more concentrated in specialties such as clinical research [27,31,35,37,48], health informatics [43,45], and research networks [25,34,40,44,49]. Collaborative research networks and clinical trials played a prominent role in the application areas.

Data sources used in the research included literature papers, institutional records, clinical documents, expert perceptions, data models, simulation models, and government databases. Technical limitations were related to performance concerns, infrastructure differences, security measures, visualization methods, and access to data sources.

Other aspects mentioned included the disparity in professionals' knowledge, the inability to process large volumes of information, and the lack of human and material resources. Legal limitations were attributed to organizational policies that restricted extensive analysis.

The main challenge reported in the studies was related to methodological approaches, particularly the inability to evaluate solutions across multiple scopes, inadequate sample sizes, limited evaluation periods, the lack of a gold standard, and the need for validation and evaluation in different study designs.

Overall, the integrated findings highlight the importance of considering the environment, application time, and methodological approaches in ensuring data quality in health research. The identified challenges and limitations provide valuable insights for future research and the development of strategies to enhance data quality assurance in various health domains.

#### **4.3.4 Strategies for Improving Data Quality**

In the analyzed studies, various strategies and interventions were used to plan, manage, and analyze the impact of implementing procedures on data quality assurance. Business intelligence models guided some studies, using extraction, transform, and load (ETL) [32,40,41,47,51]; preprocessing [28,45,52-54]; Six Sigma practices [32,48]; and the business process management (BPM) model [33]. Data monitoring strategies included risk-based approaches [36,37], data source verification [35,37,38], central monitoring [37,38], remote

monitoring (eg, telephone contact) [31,38], and training [29]. Benchmarking strategies were applied across systems or projects in some cases [26,50,51].

Quantitative analyses primarily involved combined strategies, with data triangulation often paired with statistical analyses. Data mining techniques [24], deep learning, and natural language processing [45] were also used in combination or individually in different studies. Statistics alone was the most commonly used quantitative technique. The qualitative analysis encompassed diverse approaches, with consultation with specialists [30,34,43,44,54,55], structured instruments [29,38,44,46], data set validation [41,42,56], and visual analysis [33,40,48] being prominent. Various qualitative techniques, such as interviews [27], the Delphi technique [24], feedback audit [35], grammatical rules [39], and compliance enforcement [49], were reported.

Different computational resources were used for analysis and processes. The R language (R Core Team and the R Foundation for Statistical Computing) was commonly used for planning and defining data sets, while Python and Java were mentioned in specific cases for auditing databases and error detection. Clinical and administrative software, web portals, and electronic data capture platforms (eg, Research Electronic Data Capture [REDCap], CommonCarecom, MalariaCare, Assistance Publique–Hôpitaux de Paris–Clinical Data Repository [AP-HP-CDR], Intensive Care Unit DaMa–Clinical information System [ICU-DaMa-CIS]) were used for support, decision-making, data set planning, collection, and auditing. Additional tools, such as dictionaries, data plans, quality indicators, data monitoring plans, electronic measurements (e-measures), and Microsoft Excel spreadsheets, were also used.

It is evident that a range of strategies, interventions, and computational resources were used to ensure data quality in the studies. Business intelligence models, statistical analyses, data mining techniques, and qualitative approaches played significant roles in analyzing and managing data quality. Various programming languages and software tools were used for different tasks, while electronic data capture platforms facilitated data collection and auditing. The integration of these findings highlights the diverse approaches and resources used to address data quality in the analyzed studies.

#### **4.3.5 Synthesis of Findings**

The main barriers reported related to the theme of research in the area of health data quality cite circumstances regarding use, systems, and health services. Such barriers are

influenced by technical, organizational, behavioral, and environmental factors that cover significant contexts of information systems, specific knowledge, and multidisciplinary techniques [43]. The quality of each data element in the 9 categories can be assessed by checking its adherence to institutional norms or by comparing and validating it with external sources [41]. Table 7 summarizes the main types of obstacles reported in the studies.

**Table 7 - Barriers to health data quality.**

<b>Barrier</b>	<b>Examples</b>
Technical	<ul style="list-style-type: none"> <li>• Restrictive data formats</li> <li>• Lack of metadata and standards</li> <li>• Absence of technical solutions (eg, interoperability)</li> <li>• Poor design quality (standards), development (flexibility), and evaluation (usability and complexity) of system designs</li> <li>• Lack of detailed information for specific searches</li> <li>• Terminology variations</li> <li>• Limited recovery capabilities</li> <li>• A large amount of unstructured data</li> <li>• Challenges with patient identification and matching</li> </ul>
Motivational	<ul style="list-style-type: none"> <li>• Lack of incentives to use data in decision-making</li> <li>• Lack of delegation of responsibilities</li> </ul>
Economical	<ul style="list-style-type: none"> <li>• Lack of investments in people, infrastructure, and organizational processes for collecting, storing, analyzing, and sharing data</li> </ul>
Political	<ul style="list-style-type: none"> <li>• Lack of confidence</li> <li>• Absence of restrictive guidelines and policies</li> <li>• Lack of clarity of role and data owners</li> </ul>
Legal	<ul style="list-style-type: none"> <li>• Intellectual property</li> <li>• Copyright</li> <li>• Data privacy</li> <li>• Interest conflicts</li> </ul>
Ethical	<ul style="list-style-type: none"> <li>• Purpose of data use</li> <li>• Impact on data holders</li> </ul>
Organizational	<ul style="list-style-type: none"> <li>• Organizational culture</li> <li>• Low dissemination of research activities</li> </ul>
Human Resources	<ul style="list-style-type: none"> <li>• Inadequate number of qualified and motivated personnel</li> <li>• Little or no supervision</li> <li>• Heavy workload</li> <li>• Team rotations</li> </ul>
Methodological	<ul style="list-style-type: none"> <li>• Sample size</li> <li>• Little or no training in data analysis and interpretation tools</li> </ul>

	<ul style="list-style-type: none"> <li>• Data extraction issues</li> <li>• Unfamiliarity with data quality assessment</li> <li>• Source document complexity</li> <li>• Study design</li> <li>• Measured variables (primary or secondary)</li> <li>• Data collection time</li> <li>• Encoding methods</li> <li>• Transcription errors</li> </ul>
--	---

Although many electronic records provide a dictionary of data from their sources, units of measurement were often neglected and adopted outside of established standards. Such “human errors” are inevitable, reinforcing the need for continuous quality assessment from the beginning of collection. However, some studies have tried to develop ontologies to allow the automated and reproducible calculation of data quality measures, although this strategy did not have great acceptance. For Feder [55], “The harmonized data quality assessment terminology, although not comprehensive, covers common and important aspects of the quality assessment practice.” Therefore, generating a data dictionary with its determined types and creating a data management plan are fundamental in the planning of research [28].

Both the way of collecting and the way of inputting data impact the expected result from a data set. Therefore, with a focus on minimizing data entry errors as an essential control strategy for clinical research studies, implementing intervention modes of technical barriers was presented as pre- and postanalysis [56]. The problems were caused by errors in the data source, extraction, transform, and load process or by limitations of the data entry tool. Extracting information to identify actionable insights by mining clinical documents can help answer quantitative questions derived from structured health quality research data sources [39].

Given the time and effort involved in the iterative error detection process, typical manual curation was considered insufficient. The primary sources of error included human and technological errors [35]. However, outliers identified by automated algorithms should be considered potential outliers, leaving the field specialists in charge [51]. In contrast, different and ambiguous definitions of data quality and related characteristics in emergency medical services were presented [55]. Such divergences were based on intuition, previous experiences, and evaluation purposes. Using definitions based on ontology or standardization is suggested to compare research methods and their results. The definitions and relationships between the different data quality dimensions were unclear, making the quality of comparative assessment difficult [52].

In terms of evaluation methods, similar definitions overlapped. The difference lay in the distribution comparison and validity verification, where the definition of distribution comparison was based on comparing a data element with an official external reference [54]. Meanwhile, the validity check was concerned with whether a particular value was an outlier, a value outside the normal range. The reasons for the existence of multiple evaluation practices were the heterogeneity of data sources about syntax (file format), schema (data structure models), and semantics (meaning and varied interpretations) [50]. There should be a standard set of data to deal with such inconsistencies and allow data transformation into a structure capable of interoperating with its electronic records [40].

Data standardization transforms databases from disparate sources into a standard format with shared specifications and structures. It also allows users from different institutions to share digital resources and can facilitate the merging of multicenter data and the development of federated research networks [34]. For this, 2 processes are necessary: (1) standardization of individual data elements, adhering to terminology specifications [49], and (2) standardization of the database structure through a minimum data set, which specifies where data values are located and stored in the database [50]. Improvements in electronic collection software functionality and its coding structures have also been reported to result in lower error rates [36].

In addition, it is recommended to know the study platform and access secondary data sources that can be used. In this way, transparency in the systemic dissemination of data quality with clear communication, well-defined processes, and instruments can improve the multidisciplinary cooperation that the area requires [44].

Awareness campaigns on the topic at the organizational level contributed to improving aspects of data governance. The most reported error prevention activities were the continuing education of professionals with regular training of data collectors during their studies [50]. In this sense, in-service education should promote the correct use of names formulated by structured systems to improve the consistency and accuracy of records and favor their regular auditing. Health systems that received financial incentives for their research obtained more satisfactory results regarding the degree of reliability of their data [53].

Figure 5 depicts the great diversity of elements involved in the data quality process in health research, representing the planning (precollection), development (data acquisition and monitoring), and analysis (postcollection) stages. In our findings, each phase presented a set of strategies and tools implemented to provide resources that helped the interaction between phases.

Figure 5 - Elements involved in the research data quality process. Elements involved in the data quality process. FAIR: Findability, Accessibility, Interoperability, and Reuse; ICHGCP: International Conference on Harmonization—Good Clinical Practice; ISO: International Organization for Standardization.

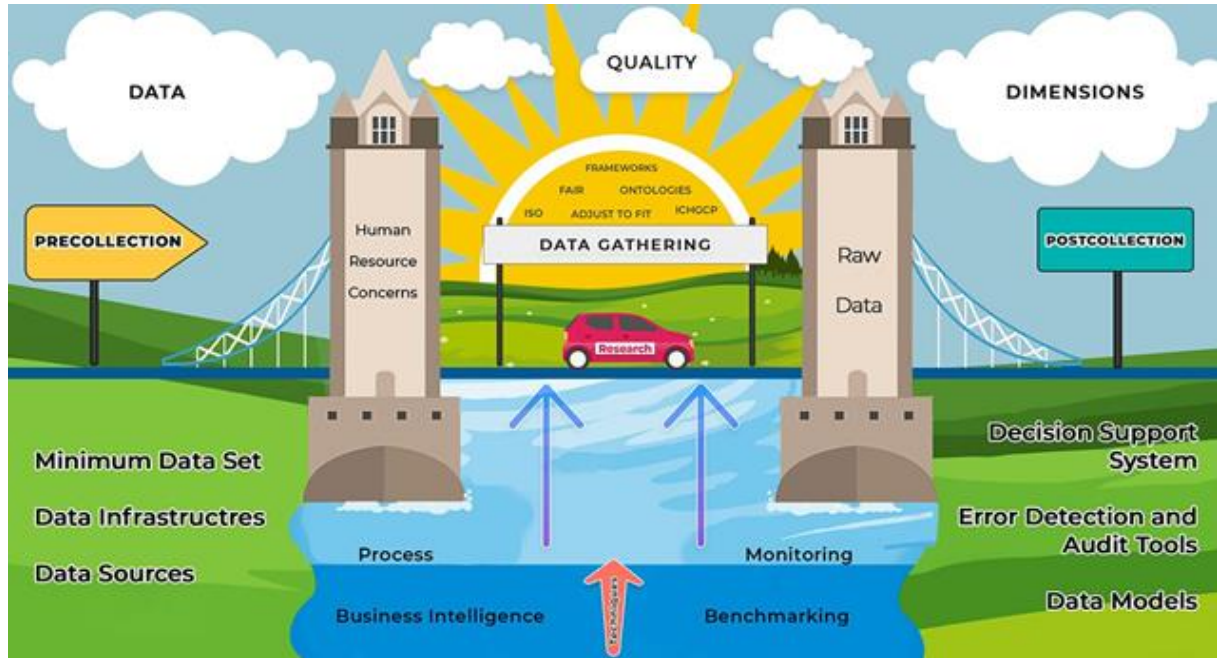


Figure 5 depicts the great diversity of elements involved in the data quality process in health research, representing the planning (precollection), development (data acquisition and monitoring), and analysis (postcollection) stages. In our findings, each phase presented a set of strategies and tools implemented to provide resources that helped the interaction between phases.

For the success of research, the processes and techniques must be fluid and applied in a direction based on good guides and recommendations. The research must go through phases, with well-established bases and tools suitable for its purpose, using sources and instruments available through digital strategies and systems, models, guides and feedback, and audit mechanisms.

In addition, every beginning of a new phase must be supported by well-defined pillars that encompass the exhaustive use of validations and pretests; plans for monitoring, management, and data analysis; precautions for ethical and legal issues; training of the team; and channels for effective communication.

In the broadest sense, incorporating data quality techniques and tools is analogous to going on a trip, that is, going from point A to point B. The starting point refers to good planning of issues, such as the year's season, the quantity and type of items that will be transported, the



most appropriate means of transport, the budget available, and tips and guidance available in the different means of communication. Even if the path is already known, an important step that precedes the beginning of its execution is always the definition of the best route. Consulting maps and updated conditions are always recommended since they can change over time.

However, the execution phase of a trip is not limited to reaching the final destination. During the journey, we should always be attentive to signs and directions, without obviously failing to enjoy the landscape and all its opportunities. Finally, when we arrive at our destination, we must bear in mind that to obtain the best results, it is necessary to know the best guides and tourist attractions. A wrong choice or decision can provide us with a low-quality photograph, an unexpected experience, and, as an effect, an epilogue of bad memories.

#### ***4.4 Discussion***

##### **4.4.1 Principal Findings**

This study presented contributions to aid the ultimate goal of good data quality focused on findings that used some digital technology (ie, to develop a disciplined process of identifying data sources, preparing the data for use, and evaluating the value of those sources for their intended use). Key findings revealed variability and a lack of consensus in assessing data quality domains and metrics. Data quality factors included the research environment, application time, and development steps. Strategies for improving data quality involved using business intelligence models, statistical analyses, data mining techniques, and qualitative approaches. The findings highlight the need for standardized practices and collaborative efforts to enhance data quality in health research.

The routine of health services that deal with demands for collecting and consuming data and information can benefit from the set of evidence on tools, processes, and evaluation techniques presented here. Increasingly ubiquitous in the daily lives of professionals, managers, and patients, technology should not be adopted without a specific purpose, as doing so can generate misinterpreted information obtained from unreliable digital health devices and systems. The resources presented can help guide medical decisions that not only involve medical professionals but also indirectly contribute to avoiding decisions based on low-quality information that can put patients' lives at risk.

With the promotion of the data culture increasingly present in a transversal way, research and researchers can offer increasingly more reliable evidence and, in this way, benefit

the promotion and approach to the health area. This mutual cycle must be transparent so that there is awareness that adherence to such a practice can favor the potential strengthening of a collaborative network based on results and promote fluidity and methodological transparency. In addition, it encourages data sharing and, consequently, the reuse of data into reliable information silos, enhancing the development and credibility of health research. At the international level, platforms with a centralized structure of reliable data repositories of patient records that offer data sharing have reduced duplication of efforts and costs. This collaboration can further decrease disparate inequities between middle- and high-income, giving celerity and minimizing risks in the development and integrity of studies.

Reliable data can play a crucial role in enlightening health institutions that prioritize cultivating a data-centric culture and are well equipped to deliver high-quality information. This, in turn, facilitates improved conditions for patient care. In addition to mapping concepts between different sources and application scenarios, it is essential to understand how initial data quality approaches are anchored in previous concepts and domains, with significant attention to suitability for use, following guidelines or using frameworks in a given context [41]. Since the concept in the same data source can change over time, it is still necessary to carry out mapping with an emphasis on its dimensions in a sensible way and on how the evolution of concepts, processes, and tools impacts the quality assessment of research and health services [47].

The realization of mapping with emphasis on domains or concepts must coexist in health information systems. The outcome favors maximizing processes, increasing productivity, reducing costs, and meeting research needs [26]. Consequently, within legal and ethical limits, it is increasingly necessary to use data comprehensively and efficiently to benefit patients [57]. For example, recent clinical and health service research has adopted the “fit for use” concept proposed in the information science literature. This concept implies that data quality dimensions do not have objective definitions but depend on tasks characterized by research methods and processes [48]. Increasingly, data quality research has borrowed concepts from various referencing disciplines. More importantly, with many different referencing disciplines using data quality as a context within their discipline, the identity of the field of research has become increasingly less distinct [33].

#### 4.4.2 Comparison With Prior Work

The large dissonance between domain definitions has increasingly motivated the search for a gold standard to be followed [30]. The area has received particular attention, especially after the term “big data” gained increasing strength [58]. The human inability to act with a large volume of information in research and the need to control this high data volume are increasingly driving the emergence of digital solutions. Although the definition of these digital data quality tools occurs from the end user’s perspective, their implementation occurs from the researcher’s perspective; a data set is highly context specific [33]. So, a generic assessment framework is unlikely to provide a comprehensive data quality analysis for a specific study, making its selection dependent on the study’s analysis plan [40].

The use of ontologies, for example, can help quantify the impact of likely problems, promote the validity of an effective electronic measure, and allow a generalization of the assessment approach to other data analysis tasks in more specific domains [55]. This benefit allows the decision-making process and planning of corrective actions and resource allocation faster [47]. However, the complex coding process can generate inconsistencies and incompleteness due to the characterization of clinically significant conditions, insufficient clinical documentation, and variability in interpretation [30]. Therefore, it is critical to use specific rules that capture relevant associations in their corresponding information groups. Administrative health data can also capture valuable information about such difficulties using standardized terminologies and monitor and compare coded data between institutions [24].

Nevertheless, as a consequence of this lack of standard, the use of integrated quality assurance methods combined with standard operating procedures (SOPs) [58], the use of rapid data feedback [38], and supportive supervision during the implementation of surveys are feasible, effective, and necessary to ensure high-quality data [31]. Adopting such well-defined interventions still plays an essential role in data quality management. It is possible to perform these activities through process control and monitoring methods, data manipulation and visualization tools, techniques, and analysis to discover patterns and perspectives on the target information subset [27]. Regardless of the model adopted, these tools should aim to discover abnormalities and provide the ability to stop and correct them in an acceptable time, also allowing for the investigation of the cause of the problem [56].

Technology is an excellent ally in these processes, and in parallel with the tools of the Lean Six Sigma philosophy, it can partially replace human work [31]. To maximize the potential of this combination, the value derived from using analytics must dictate data quality

requirements. Computer vision/deep learning, a technology to visualize multidimensional data, has demonstrated data quality checks with a systematic approach to guarantee a reliable and viable developed asset for health care organizations for the holistic implementation of machine learning processes [53]. However, most of these analytical tools still assume that the analyzed data have high intrinsic quality, which can thus allow possible failures in the process, in addition to the final experiments' lack of optimization, safety, and reliability [37].

In this way, the reuse of information has a tremendous negative impact [48]. The centralized storage of variables without excellent mapping to changes in system paradigms (metadata) and with a mechanism to trace the effects of changes in concepts that are frequent in the health area can also affect the reliability of research [37]. For example, the severity classification of a given condition can change over time and, consequently, mitigate the comparability power of a study or even prevent it from being used as a basis for planning or evaluating a new one [52]. In addition, the cultural background and experience of researchers can influence the interpretation of data [44]. Therefore, a combination of integrated tools located centrally and at each partner site for decentralized research networks can increase the quality of research data [40].

A central metadata repository contains common data elements and value definitions used to validate the content of data warehouses operated at each location [34]. So, the consortium can work with standardized reports on data quality, preserving the autonomy of each partner site and allowing individual centers to improve data in their locally sourced systems [29]. It is, therefore, essential to consider the quality of a record's content, the data quality usability, and what mechanisms can make data available for broader use [41]. As outlined by Kodra et al [42], managing data at the source and applying the FAIR guiding principles for data management are recognized as fundamental strategies in interdisciplinary research network collaboration.

Data production and quality information dissemination depend on establishing a record governance model; identifying the correct data sources; specifying data elements, case report forms, and standardization; and building an IT infrastructure per agreed principles [29]. Developing adequate documentation, training staff, and providing audit data quality are also essential and can serve as a reference for teaching material for health service education [25]. This can facilitate more quality studies in low- and middle-income countries.

The lack of such studies implies that health systems and research performance in these countries still face significant challenges at strategic stages, such as planning and managing complete data, leading to errors in population health management and clinical care [43]. In turn,

the low use of health information and poor management of health information systems in these countries make evidence-based decisions and planning at the community level difficult [2]. The results also demonstrate that, despite existing, such individual training efforts focus mainly on transmitting data analysis skills [33].

#### **4.4.3 Strengths**

Identifying systematic and persistent defects in advance and correctly directing human, technical, and financial resources are essential to promote better management and increase the quality of information and results achieved in research [42]. This step can provide improvements and benefits to health managers, allowing greater efficiency in services and better allocation of resources. Promoting such benefits to society through relevant data impacts the performance and effectiveness of public health services [39] and boosts areas of research, innovation, and enterprise development [59].

Creative approaches to decision-making in data quality and usability require good use of transdisciplinary collaboration among experts from various fields regardless of study design planning or application area [59]. This use may be reaching the threshold of significant growth and thus forcing the need for a metamorphosis from the measurement and evaluation of data quality, today focused on content, to a direction focused on use and context [57].

Without a standard definition, the use of the “fit for purpose” concept for performance monitoring, program management, and data quality decision-making is growing. As a large part of this quality depends on the collection stage, interventions must target the local level where it occurs and must encompass professionals at the operational level and forms at the technical level. Identifying and addressing behavioral and organizational challenges and building technical capacity are critical [60], increasingly fostering a data-driven culture [29,30].

#### **4.4.4 Limitations**

Among the limitations of our review, we first highlight the search for works written in English and Portuguese, since the interpretation of concepts and even the literal translations of terms referring to the dimensions and adaptations to different cultural realities can vary, and thus influenced part of our evaluation [31]. The limitation may impact the results by excluding

relevant research published in other languages and overlooking diverse cultural perspectives. To mitigate this, we suggest expanding collaboration with multilingual experts and including studies in various languages to ensure a comprehensive and unbiased evaluation of data quality.

Second, the absence of evidence in middle-income countries prevented the authors from conducting an adequate synthesis regarding the performance and application of the evidence found in these countries [2]. Limited representation from middle-income countries hinders the generalizability and applicability of findings, risking a biased understanding of intervention effectiveness. Inclusion of more studies from middle-income countries is vital for comprehensive evidence synthesis, enabling better comprehension of intervention performance in worldwide contexts and avoiding oversight of critical perspectives and outcome variations.

Third, due to the rapid growth of technologies applied to the area, we conducted a search focused on the past 5 years, which may draw attention away from other fundamentals and relevant procedures. The limited time span may lead to incomplete findings and conclusions, hindering a comprehensive understanding of the field's knowledge and advancements. To address this limitation, future research should consider a broader time frame to include older studies, allowing for a more thorough examination of fundamentals and relevant procedures impacted by the rapid evolution of technologies in the area.

#### **4.4.5 Future Directions**

Once the technical and organizational barriers have been overcome, with data managed, reused, stored, extracted, and appropriately distributed [46], health care must also pay attention to behavior focused on interactions between human, artificial, and hybrid actors. This interaction reflects the importance of adhering to social, ethical, and professional norms, including demands related to justice, responsibility, and transparency [60]. In short, increasing dependence on quality information increases its possibilities [61], but it also presents regulators and policy makers with considerable challenges related to their governance in health.

For future work, developing a toolkit based on process indicators is desirable to verify the quality of existing records and provide a score and feedback on the aspects of the registry that require improvements. There is a need for coordination between undergoing initiatives at national and international levels. At the national level, we recommend developing a centralized, public, national "registration as a service" platform, which will guarantee access to highly trained personnel on all topics mentioned in this paper, promoting the standardization of

registries. In addition to allowing cost and time savings in creating new registries, the strategy should allow for linking essential data sources on different diseases and increase the capacity to develop cooperation at the regional level.

We also suggest using the data models found in this study to serve as a structured information base for decision support information system development and health observatories, which are increasingly relevant to public health. Furthermore, concerning the health context, it may allow the execution of implementation research projects and the combination with frameworks that relate to health behavior interventions, for example, the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework [62], among others.

#### 4.4.6 Conclusion

This study will help researchers, data managers, auditors, and systems engineers think about the conception, monitoring, tools, and methodologies used to design, execute, and evaluate their research and proposals concerned with data quality. With a well-established and validated data quality workflow for health care, it is expected to assist in mapping the management processes of health care research and promote the identification of gaps in the collection flow where any necessary data quality intervention can be accordingly evaluated with the best tools described here. In conclusion, the results provide evidence of the best practices using data quality approaches involving many other stakeholders, not just researchers and research networks. Although there are some well-known data quality guidelines, they are context specific and not found in the identified scientific publications. So, the information collected in this study can support better decision-making in the area and provide insights that are distinct from the context-specific information typically found in scientific publications.

#### 4.4.7 References

1. Hekler E, Tiro JA, Hunter CM, Nebeker C. Precision health: the role of the social and behavioral sciences in advancing the vision. *Ann Behav Med* 2020 Nov 01;54(11):805-826 [[FREE Full text](#)] [doi: [10.1093/abm/kaaa018](https://doi.org/10.1093/abm/kaaa018)] [Medline: [32338719](#)]
2. Harrison K, Rahimi N, Danovaro-Holliday MC. Factors limiting data quality in the expanded programme on immunization in low and middle-income countries: a scoping review. *Vaccine* 2020 Jun 19;38(30):4652-4663 [[FREE Full text](#)] [doi: [10.1016/j.vaccine.2020.02.091](https://doi.org/10.1016/j.vaccine.2020.02.091)] [Medline: [32446834](#)]
3. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst* 2009 Mar;24(2):8-12 [doi: [10.1109/mis.2009.36](https://doi.org/10.1109/mis.2009.36)]

4. Knight W. The foundations of AI are riddled with errors. *Wired*. 2021 Mar 31. URL: <https://www.wired.com/story/foundations-ai-riddled-errors/> [accessed 2023-10-18]
5. Assaf A, Senart A. Data quality principles in the semantic web. 2012 Presented at: ICSC2012: 6th IEEE International Conference on Semantic Computing; September 19-21, 2012; Palermo, Italy [doi: [10.1109/icsc.2012.39](https://doi.org/10.1109/icsc.2012.39)]
6. Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S. Quality assessment for linked data: a survey. Maastricht University. 2017 May 5. URL: <https://cris.maastrichtuniversity.nl/en/publications/quality-assessment-for-linked-data-a-survey> [accessed 2023-06-09]
7. Peng C, Goswami P. Meaningful integration of data from heterogeneous health services and home environment based on ontology. *Sensors (Basel)* 2019 Apr 12;19(8):1747 [FREE Full text] [doi: [10.3390/s19081747](https://doi.org/10.3390/s19081747)] [Medline: [31013678](https://pubmed.ncbi.nlm.nih.gov/31013678/)]
8. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016 Sep 11;4(1):1244 [FREE Full text] [doi: [10.13063/2327-9214.1244](https://doi.org/10.13063/2327-9214.1244)] [Medline: [27713905](https://pubmed.ncbi.nlm.nih.gov/27713905/)]
9. World Health Organization. Improving Data Quality: A Guide for Developing Countries. Manila: World Health Organization. Regional Office for the Western Pacific; 2003.
10. Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. *BMC Health Serv Res* 2017 Nov 22;17(1):766 [FREE Full text] [doi: [10.1186/s12913-017-2697-y](https://doi.org/10.1186/s12913-017-2697-y)] [Medline: [29166905](https://pubmed.ncbi.nlm.nih.gov/29166905/)]
11. A experiência brasileira em sistemas de informação em saúde Internet;1. Ministério da Saúde do Brasil, Organização Pan-Americana da Saúde, Fundação Oswaldo Cruz. 2009. URL: [http://bvsmis.saude.gov.br/bvsmis/publicacoes/experiencia\\_brasileira\\_sistemas\\_saude\\_volume1.pdf](http://bvsmis.saude.gov.br/bvsmis/publicacoes/experiencia_brasileira_sistemas_saude_volume1.pdf) [accessed 2023-10-18]
12. Ndabarora E, Chipps JA, Uys L. Systematic review of health data quality management and best practices at community and district levels in LMIC. *Inf Dev* 2013 Jun 27;30(2):103-120 [doi: [10.1177/0266666913477430](https://doi.org/10.1177/0266666913477430)]
13. Mulgund P, Sharman R, Anand P, Shekhar S, Karadi P. Data quality issues with physician-rating websites: systematic review. *J Med Internet Res* 2020 Sep 28;22(9):e15916 [FREE Full text] [doi: [10.2196/15916](https://doi.org/10.2196/15916)] [Medline: [32986000](https://pubmed.ncbi.nlm.nih.gov/32986000/)]
14. Benchoufi M, Ravaud P. Blockchain technology for improving clinical research quality. *Trials* 2017 Jul 19;18(1):335 [FREE Full text] [doi: [10.1186/s13063-017-2035-z](https://doi.org/10.1186/s13063-017-2035-z)] [Medline: [28724395](https://pubmed.ncbi.nlm.nih.gov/28724395/)]
15. Lovis C. Unlocking the power of artificial intelligence and big data in medicine. *J Med Internet Res* 2019 Nov 08;21(11):e16607 [FREE Full text] [doi: [10.2196/16607](https://doi.org/10.2196/16607)] [Medline: [31702565](https://pubmed.ncbi.nlm.nih.gov/31702565/)]
16. Pellison FC, Rijo RPCL, Lima VC, Crepaldi NY, Bernardi FA, Galliez RM, et al. Data integration in the Brazilian Public Health System for tuberculosis: use of the semantic web to establish interoperability. *JMIR Med Inform* 2020 Jul 06;8(7):e17176 [FREE Full text] [doi: [10.2196/17176](https://doi.org/10.2196/17176)] [Medline: [32628611](https://pubmed.ncbi.nlm.nih.gov/32628611/)]
17. Stern C, Lizarondo L, Carrier J, Godfrey C, Rieger K, Salmond S, et al. Methodological guidance for the conduct of mixed methods systematic reviews. *JBIS Evid Synth* 2020 Oct;18(10):2108-2118 [doi: [10.1112/JBISIR-D-19-00169](https://doi.org/10.1112/JBISIR-D-19-00169)] [Medline: [32813460](https://pubmed.ncbi.nlm.nih.gov/32813460/)]
18. Aromataris E, Munn Z. JBI manual for evidence synthesis. JBI. 2020. URL: <https://jbi-global-wiki.refined.site/space/MANUAL> [accessed 2023-10-18]
19. Pellizzon RDF. Pesquisa na área da saúde: 1. Base de dados DeCS (Descritores em Ciências da Saúde). *Acta Cir Bras* 2004 Apr;19(2):153-163 [doi: [10.1590/s0102-86502004000200013](https://doi.org/10.1590/s0102-86502004000200013)]
20. Romano L. Using Medical Subject Headings (MeSH) in cataloging. *Techn Serv Q* 2018 Jan 29;35(2):217-219 [doi: [10.1080/07317131.2018.1425351](https://doi.org/10.1080/07317131.2018.1425351)]
21. Paez A. Gray literature: an important resource in systematic reviews. *J Evid Based Med* 2017 Aug;10(3):233-240 [FREE Full text] [doi: [10.1111/jebm.12266](https://doi.org/10.1111/jebm.12266)] [Medline: [28857505](https://pubmed.ncbi.nlm.nih.gov/28857505/)]
22. Mendeley. Mendeley Reference Manager [disk]. Version 2.40.0. London: Mendeley Ltd; 2022.
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg* 2021 Apr;88:105906 [FREE Full text] [doi: [10.1016/j.ijsu.2021.105906](https://doi.org/10.1016/j.ijsu.2021.105906)] [Medline: [33789826](https://pubmed.ncbi.nlm.nih.gov/33789826/)]
24. Peng M, Lee S, D'Souza AG, Doktorchik CTA, Quan H. Development and validation of data quality rules in administrative health data using association rule mining. *BMC Med Inform Decis Mak* 2020 Apr 25;20(1):75 [FREE Full text] [doi: [10.1186/s12911-020-1089-0](https://doi.org/10.1186/s12911-020-1089-0)] [Medline: [32334599](https://pubmed.ncbi.nlm.nih.gov/32334599/)]
25. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Med Res Methodol* 2021 Apr 02;21(1):63 [FREE Full text] [doi: [10.1186/s12874-021-01252-7](https://doi.org/10.1186/s12874-021-01252-7)] [Medline: [33810787](https://pubmed.ncbi.nlm.nih.gov/33810787/)]



26. Harkener S, Stausberg J, Hagel C, Siddiqui R. Towards a core set of indicators for data quality of registries. *Stud Health Technol Inform* 2019 Sep 03;267:39-45 [doi: [10.3233/SHTI190803](https://doi.org/10.3233/SHTI190803)] [Medline: [31483252](https://pubmed.ncbi.nlm.nih.gov/31483252/)]
27. Ni K, Chu H, Zeng L, Li N, Zhao Y. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. *BMJ Open* 2019 Jul 02;9(7):e029314 [FREE Full text] [doi: [10.1136/bmjopen-2019-029314](https://doi.org/10.1136/bmjopen-2019-029314)] [Medline: [31270120](https://pubmed.ncbi.nlm.nih.gov/31270120/)]
28. Huser V, Li X, Zhang Z, Jung S, Park RW, Banda J, et al. Extending Achilles Heel data quality tool with new rules informed by multi-site data quality comparison. *Stud Health Technol Inform* 2019 Aug 21;264(5):1488-1489 [doi: [10.3233/SHTI190498](https://doi.org/10.3233/SHTI190498)] [Medline: [31438195](https://pubmed.ncbi.nlm.nih.gov/31438195/)]
29. Burnett SM, Wun J, Evance I, Davis KM, Smith G, Lussiana C, et al. Introduction and evaluation of an electronic tool for improved data quality and data use during malaria case management supportive supervision. *Am J Trop Med Hyg* 2019 Apr;100(4):889-898 [FREE Full text] [doi: [10.4269/ajtmh.18-0366](https://doi.org/10.4269/ajtmh.18-0366)] [Medline: [30793695](https://pubmed.ncbi.nlm.nih.gov/30793695/)]
30. Scobie HM, Edelstein M, Nicol E, Morice A, Rahimi N, MacDonald NE, et al. SAGE Working Group on Immunization and Surveillance Data Quality and Use. Improving the quality and use of immunization and surveillance data: summary report of the Working Group of the Strategic Advisory Group of Experts on Immunization. *Vaccine* 2020 Oct 27;38(46):7183-7197 [FREE Full text] [doi: [10.1016/j.vaccine.2020.09.017](https://doi.org/10.1016/j.vaccine.2020.09.017)] [Medline: [32950304](https://pubmed.ncbi.nlm.nih.gov/32950304/)]
31. Gass JD, Misra A, Yadav MNS, Sana F, Singh C, Mankar A, et al. Implementation and results of an integrated data quality assurance protocol in a randomized controlled trial in Uttar Pradesh, India. *Trials* 2017 Sep 07;18(1):418 [FREE Full text] [doi: [10.1186/s13063-017-2159-1](https://doi.org/10.1186/s13063-017-2159-1)] [Medline: [28882167](https://pubmed.ncbi.nlm.nih.gov/28882167/)]
32. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. *Comput Methods Programs Biomed* 2019 Nov;181:104804 [doi: [10.1016/j.cmpb.2018.10.016](https://doi.org/10.1016/j.cmpb.2018.10.016)] [Medline: [30497872](https://pubmed.ncbi.nlm.nih.gov/30497872/)]
33. Andrews R, Wynn M, Vallmuur K, Ter Hofstede AHM, Bosley E, Elcock M, et al. Leveraging data quality to better prepare for process mining: an approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in Queensland. *Int J Environ Res Public Health* 2019 Mar 29;16(7):1138 [FREE Full text] [doi: [10.3390/ijerph16071138](https://doi.org/10.3390/ijerph16071138)] [Medline: [30934913](https://pubmed.ncbi.nlm.nih.gov/30934913/)]
34. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc* 2020 Dec 09;27(12):1999-2010 [FREE Full text] [doi: [10.1093/jamia/ocaa245](https://doi.org/10.1093/jamia/ocaa245)] [Medline: [33166397](https://pubmed.ncbi.nlm.nih.gov/33166397/)]
35. Houston L, Probst Y, Martin A. Assessing data quality and the variability of source data verification auditing methods in clinical research settings. *J Biomed Inform* 2018 Jul;83:25-32 [FREE Full text] [doi: [10.1016/j.jbi.2018.05.010](https://doi.org/10.1016/j.jbi.2018.05.010)] [Medline: [29783038](https://pubmed.ncbi.nlm.nih.gov/29783038/)]
36. Tian Q, Liu M, Min L, An J, Lu X, Duan H. An automated data verification approach for improving data quality in a clinical registry. *Comput Methods Programs Biomed* 2019 Nov;181:104840 [doi: [10.1016/j.cmpb.2019.01.012](https://doi.org/10.1016/j.cmpb.2019.01.012)] [Medline: [30777618](https://pubmed.ncbi.nlm.nih.gov/30777618/)]
37. Houston L, Martin A, Yu P, Probst Y. Time-consuming and expensive data quality monitoring procedures persist in clinical trials: a national survey. *Contemp Clin Trials* 2021 Apr;103:106290 [doi: [10.1016/j.cct.2021.106290](https://doi.org/10.1016/j.cct.2021.106290)] [Medline: [33503495](https://pubmed.ncbi.nlm.nih.gov/33503495/)]
38. Houston L, Probst Y, Yu P, Martin A. Exploring data quality management within clinical trials. *Appl Clin Inform* 2018 Jan 31;9(1):72-81 [FREE Full text] [doi: [10.1055/s-0037-1621702](https://doi.org/10.1055/s-0037-1621702)] [Medline: [29388180](https://pubmed.ncbi.nlm.nih.gov/29388180/)]
39. Malmasi S, Hosomura N, Chang LS, Brown CJ, Skentzos S, Turchin A. Extracting healthcare quality information from unstructured data. *AMIA Annu Symp Proc* 2017 Apr 16;2017:1243-1252 [FREE Full text] [Medline: [29854193](https://pubmed.ncbi.nlm.nih.gov/29854193/)]
40. Juárez D, Schmidt E, Stahl-Toyota S, Ückert F, Lablans M. A generic method and implementation to evaluate and improve data quality in distributed research networks. *Methods Inf Med* 2019 Sep 12;58(2-03):86-93 [FREE Full text] [doi: [10.1055/s-0039-1693685](https://doi.org/10.1055/s-0039-1693685)] [Medline: [31514209](https://pubmed.ncbi.nlm.nih.gov/31514209/)]
41. Kodra Y, Posada de la Paz M, Coi A, Santoro M, Bianchi F, Ahmed F, et al. Data quality in rare diseases registries. *Adv Exp Med Biol* 2017;1031:149-164 [doi: [10.1007/978-3-319-67144-4\\_8](https://doi.org/10.1007/978-3-319-67144-4_8)] [Medline: [29214570](https://pubmed.ncbi.nlm.nih.gov/29214570/)]
42. Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier S, van Enckevort D, et al. Recommendations for improving the quality of rare disease registries. *Int J Environ Res Public Health* 2018 Aug 03;15(8):1644 [FREE Full text] [doi: [10.3390/ijerph15081644](https://doi.org/10.3390/ijerph15081644)] [Medline: [30081484](https://pubmed.ncbi.nlm.nih.gov/30081484/)]
43. Kumar M, Gotz D, Nutley T, Smith JB. Research gaps in routine health information system design barriers to data quality and use in low- and middle-income countries: a literature review. *Int J Health Plann Manag* 2018 Jan 02;33(1):e1-e9 [doi: [10.1002/hpm.2447](https://doi.org/10.1002/hpm.2447)] [Medline: [28766742](https://pubmed.ncbi.nlm.nih.gov/28766742/)]

44. van der Bij S, Khan N, Ten Veen P, de Bakker DH, Verheij R. Improving the quality of EHR recording in primary care: a data quality feedback tool. *J Am Med Inform Assoc* 2017 Jan;24(1):81-87 [FREE Full text] [doi: [10.1093/jamia/ocw054](https://doi.org/10.1093/jamia/ocw054)] [Medline: [27274019](https://pubmed.ncbi.nlm.nih.gov/27274019/)]
45. Juddoo S, George C. Discovering most important data quality dimensions using latent semantic analysis. 2018 Presented at: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD 2018); August 6-7, 2018; Durban, South Africa [doi: [10.1109/icabcd.2018.8465129](https://doi.org/10.1109/icabcd.2018.8465129)]
46. Pozzar R, Hammer MJ, Underhill-Blazey M, Wright AA, Tulsy JA, Hong F, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. *J Med Internet Res* 2020 Oct 07;22(10):e23021 [FREE Full text] [doi: [10.2196/23021](https://doi.org/10.2196/23021)] [Medline: [33026360](https://pubmed.ncbi.nlm.nih.gov/33026360/)]
47. Sarafidis M, Tarousi M, Anastasiou A, Pitoglou S, Lampoukas E, Spetsariasis A, et al. Data quality challenges in a learning health system. *Stud Health Technol Inform* 2020 Jun 16;270:143-147 [FREE Full text] [doi: [10.3233/SHTI200139](https://doi.org/10.3233/SHTI200139)] [Medline: [32570363](https://pubmed.ncbi.nlm.nih.gov/32570363/)]
48. Shaheen NA, Manezhi B, Thomas A, AlKelya M. Reducing defects in the datasets of clinical research studies: conformance with data quality metrics. *BMC Med Res Methodol* 2019 May 10;19(1):98 [FREE Full text] [doi: [10.1186/s12874-019-0735-7](https://doi.org/10.1186/s12874-019-0735-7)] [Medline: [31077148](https://pubmed.ncbi.nlm.nih.gov/31077148/)]
49. D'Amore JD, Li C, McCrary L, Niloff JM, Sittig DF, McCoy AB, et al. Using clinical data standards to measure quality: a new approach. *Appl Clin Inform* 2018 Apr;9(2):422-431 [FREE Full text] [doi: [10.1055/s-0038-1656548](https://doi.org/10.1055/s-0038-1656548)] [Medline: [29898468](https://pubmed.ncbi.nlm.nih.gov/29898468/)]
50. Sirgo G, Esteban F, Gómez J, Moreno G, Rodríguez A, Blanch L, et al. Validation of the ICU-DaMa tool for automatically extracting variables for minimum dataset and quality indicators: the importance of data quality assessment. *Int J Med Inform* 2018 Apr;112:166-172 [doi: [10.1016/j.ijmedinf.2018.02.007](https://doi.org/10.1016/j.ijmedinf.2018.02.007)] [Medline: [29500016](https://pubmed.ncbi.nlm.nih.gov/29500016/)]
51. Sunderland KM, Beaton D, Fraser J, Kwan D, McLaughlin PM, Montero-Odasso M, ONDRI Investigators; et al. The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project. *BMC Med Res Methodol* 2019 May 15;19(1):102 [FREE Full text] [doi: [10.1186/s12874-019-0737-5](https://doi.org/10.1186/s12874-019-0737-5)] [Medline: [31092212](https://pubmed.ncbi.nlm.nih.gov/31092212/)]
52. Pezoulas V, Kourou K, Kalatzis F, Exarchos T, Venetsanopoulou A, Zampeli E, et al. Medical data quality assessment: on the development of an automated framework for medical data curation. *Comput Biol Med* 2019 Apr;107:270-283 [FREE Full text] [doi: [10.1016/j.combiomed.2019.03.001](https://doi.org/10.1016/j.combiomed.2019.03.001)] [Medline: [30878889](https://pubmed.ncbi.nlm.nih.gov/30878889/)]
53. Chang H, Huang E, Hou I, Liu H, Li F, Chiou S. Using a text mining approach to explore the recording quality of a nursing record system. *J Nurs Res* 2019 Jun;27(3):e27 [FREE Full text] [doi: [10.1097/jnr.0000000000000295](https://doi.org/10.1097/jnr.0000000000000295)] [Medline: [30694223](https://pubmed.ncbi.nlm.nih.gov/30694223/)]
54. Mashoufi M, Ayatollahi H, Khorasani-Zavareh D. A review of data quality assessment in emergency medical services. *Open Med Inform J* 2018 May 31;12(1):19-32 [FREE Full text] [doi: [10.2174/1874431101812010019](https://doi.org/10.2174/1874431101812010019)] [Medline: [29997708](https://pubmed.ncbi.nlm.nih.gov/29997708/)]
55. Feder SL. Data quality in electronic health records research: quality domains and assessment methods. *West J Nurs Res* 2018 May 24;40(5):753-766 [doi: [10.1177/0193945916689084](https://doi.org/10.1177/0193945916689084)] [Medline: [28322657](https://pubmed.ncbi.nlm.nih.gov/28322657/)]
56. Johnson S, Speedie S, Simon G, Kumar V, Westra B. Quantifying the effect of data quality on the validity of an eMeasure. *Appl Clin Inform* 2017 Dec 14;08(04):1012-1021 [doi: [10.4338/aci-2017-03-ra-0042](https://doi.org/10.4338/aci-2017-03-ra-0042)]
57. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 2019 Nov 04;1(11):501-507 [doi: [10.1038/s42256-019-0114-4](https://doi.org/10.1038/s42256-019-0114-4)]
58. Shankaranarayanan G, Blake R. From content to context: the evolution and growth of data quality research. *J Data Inf Qual* 2017 Jan 04;8(2):1-28 [doi: [10.1145/2996198](https://doi.org/10.1145/2996198)]
59. Keller S, Korkmaz G, Orr M, Schroeder A, Shipp S. The evolution of data quality: understanding the transdisciplinary origins of data quality concepts and approaches. *Annu Rev Stat Appl* 2017 Mar 07;4(1):85-108 [doi: [10.1146/annurev-statistics-060116-054114](https://doi.org/10.1146/annurev-statistics-060116-054114)]
60. Morley J, Luciano F. How to design a governable digital health ecosystem. In: *The 2020 Yearbook of the Digital Ethics Lab*. Cham: Springer; 2021:69-88
61. Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes. *Philos Technol* 2017 Aug 15;31(4):611-627 [doi: [10.1007/s13347-017-0279-x](https://doi.org/10.1007/s13347-017-0279-x)]
62. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. *Front Public Health* 2019;7:64 [FREE Full text] [doi: [10.3389/fpubh.2019.00064](https://doi.org/10.3389/fpubh.2019.00064)] [Medline: [30984733](https://pubmed.ncbi.nlm.nih.gov/30984733/)]

## 5 THE MINIMUM DATA SET FOR RARE DISEASES: SYSTEMATIC REVIEW

### *5.1 Introduction*

#### **5.1.1 Background**

The minimum data set (MDS) is a collection of data elements to be grouped in a standard approach to allow the use of data for clinical and research purposes. An MDS is designed to capture essential data elements, which can be aggregated, at an individual level, or a combination of both, depending on the specific requirements of the registry and research objectives.

Data standardization allows for the accurate comparability of collected data and, consequently, improved generalization of findings [1]. Many countries use the MDS strategy in their health systems to standardize essential and fundamental data elements to properly record patient information and support public health planning and management. In addition, MDSs facilitate data interoperability between different health services that comprise a national health network [2]. The MDS strategy makes it possible to identify relevant health indicators in health information systems (HISs) that can serve as a basis for the definition of public policies that can influence the monitoring of diseases, organize the resources available, and improve social well-being [2].

With the high volume of fragmented data in different health services, a validated, stable, and safe MDS is essential to support the use of clinical-administrative elements and information to generate national health statistics. As noted by the World Health Organization (WHO), it can help elaborate models of service performance and patient satisfaction to meet the growing health demands of HISs worldwide. Therefore, the development of HISs relevant to public health should be able not only to generate data but also understand and manage the produced information. They should also provide indicators to represent the true health context of a given population and provide subsidies to monitor the quality of treatments offered by health services [3].

However, health data are typically voluminous, complex, and sometimes ambiguous to generate indicators that can provide knowledge and information on health. Thus, raw data must be ordered, interpreted, and transformed into information [4] that must be processed and analyzed. Many digital tools are used for processing and analysis, including machine learning

algorithms, artificial intelligence, neural network applications, big data, computational ontologies, and semantic web [5].

This complexity extends further to the rare disease (RD) domain. RDs are pathologies with a low prevalence in the general population, with <50 cases per 100,000 individuals [6]. Although they affect a low percentage of people individually, such diseases are numerous, and together they can affect up to 10% of the world population; thus, they significantly impact health systems [7]. For many RDs, there is no well-structured knowledge about their diagnoses and treatments, increasing the demand for MDS strategies [8].

Many countries have initiated national plans to promote care, research, and technology in RD. These plans focus on enabling health managers to improve the services provided in a contextualized way [9]. In addition, these proposals can increase the accuracy of health decisions and reduce the fragmentation of large volumes of data, creating a solid base of information pertinent to diagnoses, treatments, and processes [10].

Owing to the complexity of the areas of knowledge related to RD, initiatives have been developed to provide informational support to health networks that provide services, care, and research in RD [11]. The best known is Orphanet, a networked platform comprising researchers from European countries. It is funded by the European Commission to increase the available knowledge base of RDs to improve care processes in this domain. Orphanet promotes and provides a structured and comprehensive database of information and knowledge about the RD domain. It also offers a validated ontology with a high standard of quality consistency and manual data auditing performed by specialists [12].

In Brazil, the Brazilian Policy for Comprehensive Care for People with Rare Diseases of the Ministry of Health of Brazil established reference services in RD that offer preventive, diagnostic, and therapeutic actions for people with rare conditions [13]. There is a need for human, technological, and infrastructure resources in the Brazilian Unified Health System, which leads to management and communication problems such as difficulties in transmitting information between the services that comprise the health network, leading to different processes and costs across the public system [14].

MDS is essential for health surveillance, providing services, and generating recommended population indicators. However, the national policy in Brazil does not support the structured use of validated MDS in health services [15]. To address this issue, the Brazilian National Network of Rare Diseases (in Portuguese: Rede Nacional de Doenças Raras [RARAS]) was designed to create an epidemiological surveillance structure. This network

encompasses all reference services for RDs enabled in Brazil and brings together university hospitals and reference services for neonatal screening from all regions of the country [16].

The RARAS project is conducting the first nationally representative survey in Brazil on the epidemiology, clinical scenarios, diagnostic and therapeutic resources, and costs related to individuals with RD of genetic and nongenetic origin. Project managers developed their own data collection protocol in the absence of a standardized global MDS reference. Although we identified this information bottleneck in Brazil, the international literature describes a global problem with data collection, recording, and structuring in RD [17].

### **5.1.2 Objectives**

This study aimed to identify and analyze the MDSs used for RD in health care networks worldwide and compare them against WHO guidelines. The secondary objectives were to verify MDS implementation and study quality, map the collection of data elements, suggest a global fundamental MDS for RD, improve diagnostic and care processes, and optimize public planning and decision-making.

## **5.2 Methods**

### **5.2.1 Research Question Definition**

The population, concept, and context (PCC) methodology proposed by the Joanna Briggs Institute was used to define the research question of this systematic review. The PCC strategy is recommended as an alternative to the population, intervention, comparison, and outcome model for investigations without well-defined clinical intervention [18]. A total of four research questions were defined, preceding our central question as follows: (1) Are there standardized MDSs for RD patient records in health networks worldwide? (2) What are the data elements of each MDS? (3) What can we assess of their usefulness? and (4) Can a fundamental MDS be developed for RD networks?

These questions were formulated to help answer the following central question: “What is the minimum data set used in registries for RDs in health networks?”

To improve the transparency and reproducibility of the review, we registered it in the Prospective Register of Systematic Reviews international database under the identifier CRD42021221593 [17].

### **5.2.2 Inclusion Criteria**

The inclusion criteria were research papers (eg, full texts and conference papers), national plans, policies, industry reports, position papers, and program reports. Studies published in Portuguese, English, and Spanish were included. Owing to the specificity and novelty of the proposal, we included publications from any time. Only studies that fully described MDS for RD and its practical implementation as a health strategy were included. Furthermore, only studies that answered our research question and provided clear evidence on the subject were included.

### **5.2.3 Exclusion Criteria**

Publications without the scientific rigor to answer our research question using clear and objective evidence were excluded from this systematic review. We excluded personal editorials, studies using only theoretical approaches without practical implementation (such as studies defining operational or predictive models), and social media publications. Studies published in languages other than those listed in the inclusion criteria were also excluded. In addition, studies that used an MDS for RD but did not describe all the data elements that comprised the MDS were excluded.

### **5.2.4 Search Strategy for Selection of Studies**

The concepts used to create the search strategy came from PCC methodology and were adapted using Medical Subject Headings, the vocabulary thesaurus of the National Library of Medicine used for indexing articles, and metadata for health sciences [19].

A total of 4 databases were reviewed: PubMed, Scopus, Web of Science, and Literatura Latino-Americana e do Caribe em Ciências da Saúde (LILACS). We also searched for documents from the WHO and government websites for additional studies. These databases were chosen because of their robustness and relevance in the clinical and health arenas. Table 8 lists the keywords and search strings entered into the search strategy for different databases. Logical connectors were used to assign the necessary precision and refine queries.

**Table 8 - Search strategies for each of the selected databases**

Database	Search strategy
PubMed	(minimum data set OR minimum data sets OR minimum data OR minimum data set OR minimum data sets OR minimum core data) AND (rare disease OR rare diseases) AND (health network OR health networks OR health service OR health services OR health administration OR public health OR health policy OR health policies)
Scopus	“minimum dataset” OR “minimum datasets” OR “minimum data” OR “minimum data set” OR “minimum data sets” OR minimum core data AND “rare disease” OR “rare diseases” AND “health network” OR “health networks” OR “health service” OR “health services” OR “health administration” OR “public health” OR “health policy” OR “health policies”
Web of science	([ALL=((minimum dataset OR minimum datasets OR minimum data OR minimum data set OR minimum data sets OR minimum core data))] AND ALL=((rare disease OR rare diseases))) AND ALL=((health network OR health networks OR health service OR health services OR health administration OR public health OR health policy OR health policies))
LILACS <sup>a</sup>	(Minimum data set OR minimum data sets OR minimum data OR minimum data set OR minimum data sets OR minimum core data) AND (rare disease OR rare diseases) AND (health network OR health networks OR health service OR health services OR health administration OR public health OR health policy OR health policies)

<sup>a</sup>LILACS: *Literatura Latino-Americana e do Caribe em Ciências da Saúde*.

The Google Scholar database was also used for manual searches and searching for references to theses and dissertations. These documents are part of the gray literature because they are not published in commercial media. Searching through such sources can reduce publication bias, increase the comprehensiveness and timeliness of reviews, and provide a balanced picture of available information [20]. RD clinical experts participated in the evaluation process.

## 5.2.5 Data Collection Processes

### 4.2.5.1 Selection of Studies

After conducting searches using the terms specified for the mentioned databases, the review authors independently screened the titles and abstracts of each retrieved article for eligibility for synthesis according to the inclusion and exclusion criteria. Next, the full texts were retrieved, and the investigators independently performed another round of reviews to determine whether these full texts met the eligibility criteria. The reviewers were not blinded to the journal titles, study authors, or associated institutions.

The points of divergence between the 2 reviewers were identified, and a third independent evaluator resolved the disparities. Search results and the process used to screen for eligibility in this phase were reported using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [21] methodology to present the number of studies considered in each review process step. All resulting studies were organized in a web service environment, and duplicates were removed. We also documented the reasons for the exclusion of each article. Subsequently, we extracted the data and mapped the general characteristics and contexts of the included studies.

### 4.2.5.2 Data Extraction

We initially conducted a pilot test to extract data from the selected articles. Subsequently, we aligned our data extraction method with the PRISMA checklist [22]. A series of information to be mapped from the selected articles was defined to cover all the details of the chosen methodology.

The general characteristics and context of the selected studies were extracted and organized into tables: authors, publication year, title, study design, territorial dimension, national context, study objective or objectives, sample size, data analysis method, study population, guidelines followed, the health domain in which the MDS was deployed, and the main findings of the selected studies.

After defining the articles included and extracting their general characteristics, we analyzed these documents to identify the MDSs used in health networks. We extracted all data elements that formed each of the implemented MDSs. To do this in an organized manner, we classified the informational elements into 10 categories: eligibility, identification, diagnosis, treatment, medical consultation, comorbidity, hospitalization, examination, outcome, and



others. These categories were determined according to WHO digital health guidelines [23] and practical projects underway, such as RARAS [8,16].

We organized all identified MDS data elements from each included study into one of these categories. We observed that many articles described MDS data elements with different names corresponding to the same information (ie, elements that used different terms but had the same meaning). For example, the data element that represents the information about the age at which the first symptoms of the disease were identified in diagnosed patients were described in different forms in different papers, such as “Age at onset,” “Age at first symptoms in clinically diagnosed patients,” or “Age at initial symptoms.” Clinical experts in RD identified these ambiguities, and different terms with the same meaning were aggregated as unique terms (the most commonly used), synonyms were recorded, and analyses of the terms’ frequencies were adjusted.

#### ***4.2.5.3 Data Categorization***

We initially compiled and categorized all the findings from the primary studies included in the systematic review to develop the base structure for fundamental MDS. This preliminary categorization followed WHO guidelines and recommendations for health indicators [3,23]. Subsequently, we held structured meetings with health professionals working in RD services in the public health system and specialist researchers on data standards and terminology for health and digital health. We combined and summarized the results of these meetings through synthesis methods with the findings obtained from the selected studies to generate the fundamental MDS proposed for the RD domain.

#### ***4.2.5.4 Risk of Bias and Quality Assessment***

Several guidelines for reporting clinical research results are available on the Equator Network website [24]. None of these recommendations focus on writing the results from a data quality perspective. The Cochrane Collaboration states that the quality assessment of the published evidence must consider the reporting of each original paper. Therefore, the quality assessment of individual studies’ findings was conducted using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist [25].

The last STROBE recommendation includes 2 topics directly relevant to data quality reporting. The first refers to the importance of measuring each data element of interest and providing data sources and details of the assessment methods, describing the comparability of

the assessment methods when applicable. The second section explains how to address and treat the missing data. In addition, qualitative evaluations of the evidence were independently performed by 2 researchers. If disagreements occurred, the issue was referred to a third researcher for a decision [26].

We adopted the STROBE recommendations to specifically cover the normalization of the importance of each data set found in the reported articles and their evaluation methods regardless of study design. After normalization into data set categories and study designs, we also qualitatively compared descriptions of their features and how to overcome and map the gaps reported in these findings.

In addition, we have used the PRISMA methodology to organize and formally present the overall results obtained by comparing individual reports and to provide a transparent evaluation according to the topics described by the PRISMA checklist. These checklist items are essential for the transparent reporting of a systematic review. We addressed most of these items, except topics related to meta-analysis, which did not apply to this review [22].

### **5.2.6 Synthesis Methods**

We aimed to combine the findings of similar studies into subcategories using a consensus among specialists. We did not exclude studies with inadequate quality from the data synthesis. Descriptive statistics were used to report the frequency and proportion of the outcome measures.

The subgroup analysis method was used to describe the possible reasons for the heterogeneity of the data extracted from the selected studies and to facilitate the synthesis of information via tabulation and visual arrangement so that we could later carry out the necessary analyses.

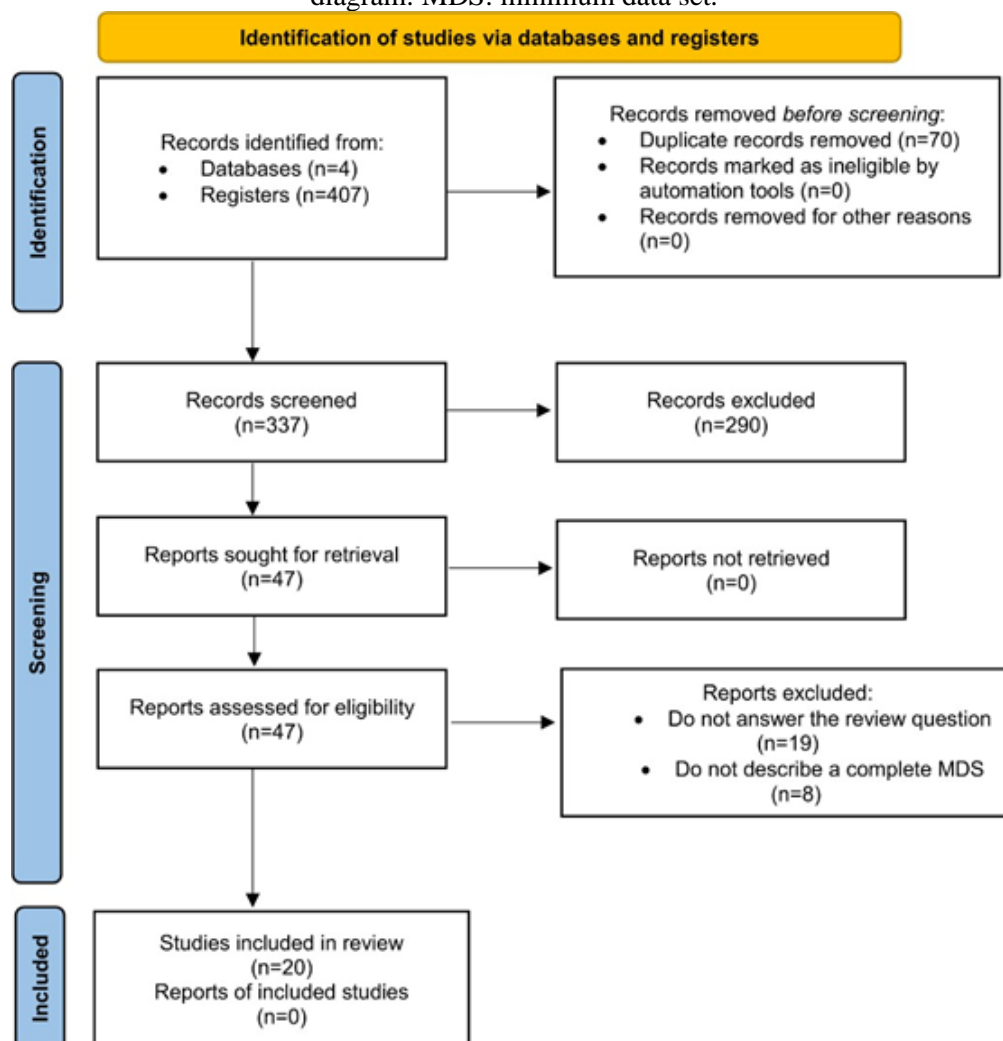
When considering the reporting method of a single instrument, proportions indicating the percentage of studies using that instrument were calculated. Next, the selected resources were analyzed and interpreted. The main contents related to the research objective were classified as RD minimum data elements. The sections were then classified into 2 general categories, management data and clinical data, which is an efficient method for categorizing health data [27].

## 5.3 Results

### 5.3.1 Description of Selected Studies

We identified 407 studies in the initial database search phase. After removing duplicates, 337 articles remained for the title and abstract screening stage, 290 of which were excluded because they were irrelevant to this review. We evaluated 47 full-text articles based on the inclusion criteria. Of these, 19 articles were excluded because they did not answer the stipulated research question, and 8 articles were excluded because they only mentioned an MDS strategy for RD and did not describe the data elements or provide information about implementation in practice. The detailed process is shown in Figure 6.

Figure 6 - PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. MDS: minimum data set.



A total of 20 unique studies were included in this systematic review. The general characteristics of these studies are described in Table 9, organized by the first author's given

names in alphabetical order. The specific characteristics of these studies are described in Multimedia Appendix 1.

**Table 9** - General characteristics of selected studies.

Study, year	Study title	Study design	Territorial dimension
Berger et al [28], 2021	How to design a registry for undiagnosed patients in the framework of rare disease diagnosis: suggestions on software, data set and coding system	Applied research	Europe
Mukhina et al [29], 2020	Primary immunodeficiencies in Russia: data from the National Registry	Cohort	Russia
Licht et al [30], 2015	The global aHUS registry: methodology and initial patient characteristics	Cohort	International
Messiaen et al [31], 2008	CEMARA: a web dynamic application within a N-tier architecture for rare diseases	Case study	France
Messiaen et al [32], 2021	10 y of CEMARA database in the AnDDI-Rares network: a unique resource facilitating research and epidemiology in developmental disorders in France	Cross-sectional cohort study	France
Stanimirovic et al [33], 2019	Development of a pilot rare disease registry: A focus group study of initial steps toward the establishment of a rare disease ecosystem in Slovenia	Case study	Slovenia
Pastores et al [34], 2007	The MPS I registry: design, methodology, and early findings of a global disease registry for monitoring patients with mucopolysaccharidosis type I	Cohort	International
Stirnadel-Farrant et al [35], 2018	Gene therapy in rare diseases: the benefits and challenges of developing a patient-centric registry for Strimvelis in ADA-SCID	Case study	Italy
Tingley et al [36], 2020	Evaluation of the quality of clinical data collection for a pan-Canadian cohort of children affected by inherited metabolic diseases: lessons learned from the Canadian Inherited Metabolic Diseases Research Network	Cohort	Canada
Subirats et al [37], 2020	Biomedical holistic ontology for people with rare diseases	Applied research	Europe
Kalankesh et al [38], 2015	Minimum data set for cystic fibrosis registry: A case study in Iran	Case study	Iran
Shahmoradi et al [39], 2019	Instructional design, development and evaluation of congenital hypothyroidism registry system	Applied research	Iran

McCann et al [40], 2014	Developing a provisional, international minimal data set for Juvenile dermatomyositis: for use in clinical practice to inform research	Applied research	The United Kingdom, Italy, and Canada
McCann et al [41], 2015	Development of an internationally agreed minimal data set for JDM for clinical and research use	Applied research	International
Huemer et al [42], 2018	Phenotype, treatment practice and outcome in the cobalamin-dependent remethylation disorders and MTHFR deficiency: data from the E-HOD registry	Cohort	Europe
Choquet et al [2], 2015	A methodology for a minimum data set for rare diseases to support national centers of excellence for health care and research	Cross-sectional	France
Derayeh et al [44], 2018	National information system for rare diseases with an approach to data architecture: A systematic review	Systematic review	Europe, the United States, Australia, and Asia
Taruscio et al [45], 2015	National registries of rare diseases in Europe: an overview of the current situation and experiences	Case study	Europe
Opladen et al [46], 2021	U-IMD: the first Unified European registry for inherited metabolic diseases	Applied research	Europe
Meissner et al [47], 2021	EULAR recommendations for a core data set for pregnancy registries in rheumatology	Applied research	Europe

### 5.3.2 Domains of Health

Of the 20 selected studies, 70% (n=14) referred to a specific health domain and the remaining 30% (n=6) referred to RD in general. Of the studies that referred to a specific domain, 2 studies were dedicated to the composition of MDS for juvenile dermatomyositis [40,41], 2 to metabolic disorders [36,45], 1 to nonmalignant RD [33], and 1 to undiagnosed RDs [28]. Other specific studies included mucopolysaccharidosis type I [34], homocystinuria and methylation defects [42], primary immunodeficiencies [29], cystic fibrosis [38], congenital hypothyroidism [39], adenosine deaminase severe combined immunodeficiency [35], atypical hemolytic uremic syndrome [30], and inflammatory rheumatic diseases [46], with 1 study each.

The remaining 6 studies that presented MDSs focused on the domain of RD in general [2,31,32,37,43,44], that is, they used data elements capable of representing patients with any RD in its health context. However, even these studies have demonstrated limitations and restrictions because of their geographic coverage and the particular local characteristics of the regions in which they were applied. As mentioned in the Discussion section, the proposed fundamental MDS seeks to address these limitations.

### 5.3.3 Description of MDSs in the Selected Studies

All the data collected and compiled, as well as their occurrences in the MDSs of the studies included in this systematic review, have been described in Multimedia Appendix 2.

In the eligibility category, the term “patient consent” appeared the most; of the 17 terms in this category, 13 were unique. In the identification category, “patient’s name” was the most common term, and 5 unique terms were observed among the 25 included in this group. Of the 148 terms in the diagnostic category, “diagnosis” was the most frequent (appearing 11 times), and 32 were unique. In the treatment category, “date treatment started” was the most common term, appearing 6 times; of the 72 terms in this category, 29 were unique. In the medical consultation category, “anthropometric data” was the term that appeared the most; of the 180 terms, 78 were unique.

Of the 17 terms, “disease malignancy” was the most common in the comorbidity category, with 8 unique terms. In hospitalization, of the 5 terms, “history of hospitalization” appeared most frequently (3 times). In the examination category, “general and specialized laboratory tests” were the most common among the 42 terms, of which 29 were unique. In the outcome category, “date of death” was the term that appeared the most, comprising 11 of the 57 terms, and in the others category, “the patient having previously provided a biological sample for research” was the most frequent term, comprising 2 of the 26 terms used in this category.

By combining these findings with the extraction, categorization, and synthesis techniques mentioned earlier, we used data science methods and clinical experience to design, structure, and recommend a fundamental global MDS for RD patient records in health care networks. It aims to comprehensively cover the data needed in clinical and management contexts. These summarized results can be found in Multimedia Appendix 3.

### 5.3.4 Risk of Bias in Included Studies

The results of the risk of bias evaluations for the 20 included studies are presented in Multimedia Appendix 4. We reported the risk of bias evaluation using the STROBE statement standards and checklist [25]. The recommendations regarding the numerical indices can be found in tab 2 of Multimedia Appendix 4, STROBE statement recommendations. We used the

“Unclear risk of bias” classification for cases where the recommendation did not apply to the study.

Using a matrix of 440 cells generated by crossing the 22 STROBE recommendations with the 20 selected studies, we found that 48.4% (213/440) of the items evaluated in all studies showed a low risk of bias and 39.1% (172/440) of these items had an unclear risk of bias. The remaining 12.5% (55/440) of the items assessed across all the studies had a high risk of bias.

In addition, the STROBE recommendations with the highest number of low risk of bias assessments for this set of studies were item 5, “Setting,” and item 14, “Descriptive data,” with 17 occurrences. The recommendation with the highest number of unclear risk of bias assessments was item 12, “Statistical methods,” with 14 occurrences. The STROBE recommendation that received the highest number of high-risk bias assessments was item 3, “Objectives,” with 12 occurrences.

## ***5.4 Discussion***

### **5.4.1 Principal Findings**

We carried out this systematic review to elucidate the MDSs used for RD in health care network records in health care systems worldwide. The results showed a lack of terminological standardization of the concepts used in these MDSs. Different health systems may use different terms for the same concepts, hindering the interoperability, integration, and sharing of highly relevant information and knowledge to describe phenomena and generate health indicators.

The WHO recommends that health systems and networks use standardized terminology to exchange data, information, knowledge, and intelligence in health. Thus, health observatories that integrate population and epidemiological data require standard semantics to collect and organize this information to enable the generation of indicators capable of positively influencing public health policies [47]. Orphanet’s initiative also points to the need for a formal vocabulary to map and classify processes in RD. The lack of structured knowledge about RDs leads to problems, such as increased waiting time for a diagnosis and, consequently, difficulties in establishing an adequate treatment for these patients [48,49].

We encountered difficulties in the standardization and classification of the MDSs findings. Notably, the selected studies did not follow the same standard structure for classifying the data from their MDSs. Thus, we classified all data into 10 categories.

This process allowed us to summarize the information, verify its frequencies, perform statistical analysis, and analyze and identify synonyms (items with the same meaning, but described using different terms in specific MDSs). Clinical experts in RD and information and data scientists met in groups to verify and analyze these synonyms.

Other studies have proposed MDSs for specific groups of RDs [28-30,34-36,38-42,45,46], populations, and territories [2,29,33,43,44] based on different methodologies. For example, previous studies have proposed MDSs for RD in national and continental contexts. The European Commission's Common Data Elements for Rare Disease Registration was developed for use by the RD registries across Europe. Similarly, the French national MDS was built nationally and defined through a national consensus for use in French RD centers [2]. However, to our knowledge, this is the first study to propose a data set for RD based on a systematic review that mapped MDSs used in different studies and could allow the application of the resulting data set to all RDs in unspecified populations.

Europe also has some examples of registries developed nationally and applied to an international consortium, such as the Fabry Disease Registry [50] and REGISTRY (for Huntington disease) [51]. However, neither provides open access to information regarding the data elements and structural evaluation. This limits researchers and practitioners who are not consortium members to access and adapt the registry to their context. In addition, these registries may limit the generalizability of the results to the entire affected population, because registered patients tend to have more access to medical care and treatment centers than the general population, especially in rural regions or countries with less developed health systems.

The Italian National Rare Diseases Registry [52] and Italian Neuromuscular Registry [53] are nationally developed registries intended to collect information on the prevalence and geographic distribution of rare and neuromuscular disorders, respectively. Both are considered effective models; however, their representativeness and generalizability of their data to other populations and countries may need to be improved.

A Spanish study addressed the importance of linking data to strengthen national RD registries [54]. However, the results were based on only 1 RD (amyotrophic lateral sclerosis, a motor neuron disease) and cannot be generalized to other diseases or datasets. In addition, this study highlights that coding errors and other inconsistencies may have affected the validity of the results. This is an important observation, because the accuracy of the results depends on the quality and consistency of the data. To address these limitations, researchers may adopt measures such as cross-checking data, manually reviewing selected cases, and statistical



analysis. This evaluation was possible because the authors performed an association of records from different data sources, which occurred exclusively in this study.

Other studies have developed MDSs for RD designed for international applications; however, the MDSs were determined by a national committee. This is the case for the National Institutes of Health, National Center for Advancing Translational Sciences, and Global Rare Diseases Patient Registry Data common data elements [55]. All these studies present MDSs with limitations or restrictions because they cover only a specific RD, or their scope of application is restricted to a specific context or geographic region.

Therefore, the most innovative aspect of this study is the compilation of all the data used in MDSs for RD published worldwide. The fundamental MDS developed based on the analysis of this compilation by clinical and health data specialists can benefit services, assistance, research, and management in RD. Most national health systems and RD policies can also contribute to developing methods and processes and producing information. Thus, defining an MDS may improve data reliability to align strategies to enhance and manage health planning [11]. Although many policies describe comprehensive care in a network without a structured and standardized MDS, health centers cannot work in a coordinated manner.

Thus, the literature indicates the need to identify and implement common data elements [56] to improve care quality [57] and enable collaboration across different health care systems [58]. Our findings can help identify common standards and data elements to minimize the duplication of efforts and enhance the quality of patient records and data. This would lead to greater effectiveness of HISs and consequently, better patient outcomes. One of the recommendations for improving the quality of RD registries is to facilitate harmonization among the many institutions that collect patient information. Thus, it is essential to plan, design, and set up an MDS-based national information system and database for RDs that can provide and evaluate health indicators, promote network research, and foster public policies for RD care [11].

MDS is usually developed by reviewing the scientific literature, consulting with experts, and considering existing guidelines and recommendations. Although some of our findings report that they followed these processes in a combined or individual way, all were performed considering only a specific geographic context or addressed aspects related to a condition or subgroup of RD. Although a systematic review is an approach to synthesizing comprehensive scientific evidence, developing MDS involves reviewing the scientific literature, consulting with experts, and considering existing guidelines to identify essential data for RDs. Our

systematic review approach provided a rigorous and transparent methodology to ensure the reliability and validity of the information collected and synthesized.

Identifying common trends and patterns can further simplify and streamline the information-gathering process, avoiding the need to develop and implement different data sets for each RD. This mapping saves time and resources, allowing for more efficient data collection. Sharing resources and knowledge plays a key role in supporting the research and development of therapies for rare conditions.

Variations in documentation practices in health systems can also lead to inconsistencies in data collection and standardization related to RD. These variations may include differences in the terminology used, coding systems, and data recording policies. Gaps in capturing these essential data elements can arise because of a lack of consensus on the most relevant information for understanding and studying RD. Existing data sets do not encompass all the aspects necessary for a comprehensive investigation, leading to the need for an MDS that identifies and incorporates essential data elements for scientific research in this field. This will allow for better comparison between different studies and registries.

Our study has some potential limitations. Among these, we highlight the methodological aspects of choosing the quality assessment method, publications, and generalization bias. In addition, we emphasize the potential for further investigations and possible outcome variations in different populations, contexts, and types of RD.

Owing to the originality and nature of the proposal, we did not define restrictions regarding the study design in our inclusion and exclusion criteria. Thus, all study designs were considered for inclusion in this systematic review. Therefore, selecting a method for assessing the quality of these studies is a nontrivial task, as such methods are usually designed to assess specific types of study designs. To mitigate this limitation, it was necessary to adopt the STROBE method in our proposal so that we could assess the quality of each of the selected studies using a standard tool. In addition, in the sequence, we cite possible biases and describe how to mitigate them.

This review may be affected by publication bias, as studies with positive results are more likely to be published than those with negative or inconclusive results. This can lead to a distorted view of the existing MDSs for RD. Therefore, we took several steps to reduce the risk of publication bias affecting our results, such as searching for multiple sources, recording the review, and contacting specialists. Recording the review process increases transparency, reduces the risk of publication bias, and promotes a more rigorous and impartial evaluation of available evidence. This helps avoid omitting relevant studies or data that may not align with

the desired results. By keeping a comprehensive record, reviewers and readers can assess the potential impact of any publication bias and ensure that the review is conducted objectively and impartially [59].

Measures adopted to minimize the risk of excluding studies that were not published in scientific journals included the use of gray literature databases, prepublication repositories, clinical trial records, and scientific conferences. Finally, we contacted experts in the field to identify unpublished or ongoing studies that were not found in the systematic search.

Possible biases that may affect the generalizability and comparability of the findings must also be mentioned. Although there was no specific tool to mitigate the issue of generalization bias, it is worth highlighting that the studies included in this analysis were selected based on clear and well-defined criteria. During the data analysis process, we considered the differences between studies, such as population characteristics, study design, and data quality. Ultimately, our goal was to present the findings of our work clearly and transparently using valid instruments from the literature (Multimedia Appendix 5). We also aimed to highlight the limitations of our study and identify potential variations in the results based on robust evidence to provide valuable and relevant information for clinical practice and health decision-making.

#### **5.4.2 Conclusions**

Our work discusses the difficulties in standardizing and classifying findings from MDSs for RD because of differences between studies. Clinical experts and data scientists have defined categories based on WHO guidelines to address this issue. This study aimed to compile MDS data from around the world and suggest a fundamental MDS for use in RD services, assistance, research, and management. The fundamental RD MDS designed in this study comprehensively covers the data needs in the clinical and management sectors.

The results can also help public policy makers achieve other aspects of their policies. For example, analyzing the state data produced by HISs is essential for qualifying and quantifying the care provided to people with RDs. It can also allow comparisons with local data regarding preventive actions provided to a community. This can improve decision-making at the managerial and local levels and contribute data that can inform strategic decisions at the national level [60].

In conclusion, the solid base of information compiled regarding MDSs for RD is a technical and social contribution to improving the health network's ability to map its demands and better understand the public health scenario regarding rare conditions. Although the proposal of a fundamental MDS for RDs is highly relevant and, to our knowledge, unprecedented in the literature, we also suggest collecting data elements to be used in addition to this fundamental MDS, if necessary, for each group or type of RD, to increase the completeness and specificity of the data structure.

Owing to the high complexity of care processes involving RDs, structured information can significantly impact the quality of services offered to the population. A curated description of the methodology for developing an MDS for RD in low- and middle-income countries has not yet been published. We encourage further research in this context.

National data gathering for RD based on standard data sets to encourage interoperability by disseminating agreed-upon data-sharing guidelines can facilitate semantic data standardization. On an organizational level, it can assist institutions in establishing a registry, sharing deidentified data with research networks, and building specific and rich databases. On the basis of these results and the proposal of a fundamental MDS, we aimed to provide evidence-based subsidies to assist managerial and clinical RD processes in health systems.

## 5.5 References

1. Shanbehzadeh M, Kazemi-Arpanahi H. Development of minimal basic data set to report COVID-19. *Med J Islam Repub Iran* 2020;34:111 [FREE Full text] [doi: [10.34171/mjiri.34.111](https://doi.org/10.34171/mjiri.34.111)] [Medline: [33315989](https://pubmed.ncbi.nlm.nih.gov/33315989/)]
2. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. *J Am Med Inform Assoc* 2015;22(1):76-85 [doi: [10.1136/amiajnl-2014-002794](https://doi.org/10.1136/amiajnl-2014-002794)]
3. Global reference list of 100 core health indicators. World Health Organization. 2015. URL: [https://apps.who.int/iris/bitstream/handle/10665/173589/WHO\\_HIS\\_HSI\\_2015.3\\_eng.pdf?sequence=1](https://apps.who.int/iris/bitstream/handle/10665/173589/WHO_HIS_HSI_2015.3_eng.pdf?sequence=1) [accessed 2022-06-07]
4. Hu Y, Miao X, Si Y, Pan E, Zio E. Prognostics and health management: a review from the perspectives of design, development and decision. *Reliab Eng Syst Saf* 2022 Jan;217:108063 [doi: [10.1016/j.ress.2021.108063](https://doi.org/10.1016/j.ress.2021.108063)]
5. Yamada DB, Bernardi FA, Miyoshi NS, de Lima IB, Vinci AL, Yoshiura VT, et al. Ontology-based inference for supporting clinical decisions in mental health. In: *Computational Science – ICCS 2020*. Cham: Springer; 2020.
6. Política Nacional de Atenção Integral às Pessoas com Doenças Raras - PNAIPDR. Ministry of Health, Government of Brazil. 2021 Jul 13. URL: <https://www.gov.br/saude/pt-br/composicao/sgtes/educamunicacao-em-doencas-raras/pnaipdr> [accessed 2023-05-23]

7. Dawkins HJ, Draghia-Akli R, Lasko P, Lau LP, Jonker AH, Cuttillo CM, International Rare Diseases Research Consortium (IRDiRC). Progress in rare diseases research 2010-2016: an IRDiRC perspective. *Clin Transl Sci* 2018 Jan 23;11(1):11-20 [[FREE Full text](#)] [doi: [10.1111/cts.12501](https://doi.org/10.1111/cts.12501)] [Medline: [28796411](https://pubmed.ncbi.nlm.nih.gov/28796411/)]
8. Félix TM, de Oliveira BM, Artifon M, Carvalho I, Bernardi FA, Schwartz IV, RARAS Network group. Epidemiology of rare diseases in Brazil: protocol of the Brazilian Rare Diseases Network (RARAS-BRDN). *Orphanet J Rare Dis* 2022 Feb 24;17(1):84 [[FREE Full text](#)] [doi: [10.1186/s13023-022-02254-4](https://doi.org/10.1186/s13023-022-02254-4)] [Medline: [35209917](https://pubmed.ncbi.nlm.nih.gov/35209917/)]
9. Bodí M, Claverias L, Esteban F, Sirgo G, De Haro L, Guardiola JJ, et al. Automatic generation of minimum dataset and quality indicators from data collected routinely by the clinical information system in an intensive care unit. *Int J Med Inform* 2021 Jan;145:104327 [doi: [10.1016/j.ijmedinf.2020.104327](https://doi.org/10.1016/j.ijmedinf.2020.104327)] [Medline: [33220573](https://pubmed.ncbi.nlm.nih.gov/33220573/)]
10. Bellgard MI, Snelling T, McGree JM. RD-RAP: beyond rare disease patient registries, devising a comprehensive data and analytic framework. *Orphanet J Rare Dis* 2019 Jul 12;14(1):176 [[FREE Full text](#)] [doi: [10.1186/s13023-019-1139-9](https://doi.org/10.1186/s13023-019-1139-9)] [Medline: [31300021](https://pubmed.ncbi.nlm.nih.gov/31300021/)]
11. Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier SL, van Enckevort D, et al. Recommendations for improving the quality of rare disease registries. *Int J Environ Res Public Health* 2018 Aug 03;15(8):1644 [[FREE Full text](#)] [doi: [10.3390/ijerph15081644](https://doi.org/10.3390/ijerph15081644)] [Medline: [30081484](https://pubmed.ncbi.nlm.nih.gov/30081484/)]
12. Procedures: Orphanet inventory of rare diseases. ORPHANET. 2017 Apr. URL: [https://www.orpha.net/orphacom/cahiers/docs/GB/eproc\\_disease\\_inventory\\_PR\\_R1\\_Nom\\_04.pdf](https://www.orpha.net/orphacom/cahiers/docs/GB/eproc_disease_inventory_PR_R1_Nom_04.pdf) [accessed 2023-04-24]
13. Portaria Nº 199, DE 30 De Janeiro De 2014. Ministry of Health, Government of Brazil. 2014. URL: [https://bvsm.s.saude.gov.br/bvs/saudelegis/gm/2014/prt0199\\_30\\_01\\_2014.html](https://bvsm.s.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html) [accessed 2023-07-10]
14. Yamada DB, Bernardi FA, Filho ME, Neiva MB, Lima VC, Vinci AL, et al. National network for rare diseases in Brazil: the computational infrastructure and preliminary results. In: Proceedings of the Computational Science- ICCS 2022: 22nd International Conference. 2022 Presented at: ICCS '22; June 21, 2022; London, UK p. 43-49 URL: [https://dl.acm.org/doi/abs/10.1007/978-3-031-08757-8\\_4](https://dl.acm.org/doi/abs/10.1007/978-3-031-08757-8_4) [doi: [10.1007/978-3-031-08757-8\\_4](https://doi.org/10.1007/978-3-031-08757-8_4)]
15. Iriart JA, Nucci MF, Muniz TP, Viana GB, de Araújo Aureliano W, Gibbon S. From the search for diagnosis to treatment uncertainties: challenges of care for rare genetic diseases in Brazil. *Cien Saude Colet* 2019;24(10):3637-3650 [[FREE Full text](#)] [doi: [10.1590/1413-812320182410.01612019](https://doi.org/10.1590/1413-812320182410.01612019)] [Medline: [31576994](https://pubmed.ncbi.nlm.nih.gov/31576994/)]
16. Alves D, Yamada DB, Bernardi FA, Carvalho I, Filho ME, Neiva MB, et al. Mapping, infrastructure, and data analysis for the Brazilian network of rare diseases: protocol for the RARASnet observational cohort study. *JMIR Res Protoc* 2021 Jan 22;10(1):e24826 [[FREE Full text](#)] [doi: [10.2196/24826](https://doi.org/10.2196/24826)] [Medline: [33480849](https://pubmed.ncbi.nlm.nih.gov/33480849/)]
17. Bernardi FA, Yamada DB, de Oliveira BM, Yamada DB, Félix TM, Alves D. The minimum dataset for rare diseases in Brazil: a systematic review protocol. *Procedia Comput Sci* 2022;196:439-444 [[FREE Full text](#)] [doi: [10.1016/j.procs.2021.12.034](https://doi.org/10.1016/j.procs.2021.12.034)]
18. The Joanna Briggs Institute reviewers' manual 2015 methodology for JBI scoping reviews. The Joanna Briggs Institute. 2015. URL: <https://reben.com.br/revista/wp-content/uploads/2020/10/Scoping.pdf> [accessed 2022-07-12]

19. Gasper W, Parvathi C, Dario G. MeSH indexing using the biomedical citation network. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2020 Presented at: BCB '20: 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; Sep 21-24, 2020; Virtual Event USA [doi: [10.1145/3388440.3412466](https://doi.org/10.1145/3388440.3412466)]
20. Paez A. Gray literature: an important resource in systematic reviews. *J Evid Based Med* 2017 Aug 31;10(3):233-240 [doi: [10.1111/jebm.12266](https://doi.org/10.1111/jebm.12266)] [Medline: [28857505](https://pubmed.ncbi.nlm.nih.gov/28857505/)]
21. PRISMA statement. PRISMA - Transparent Reporting of Systematic Reviews and Meta-Analyses. 2021. URL: <http://www.prisma-statement.org/PRISMAStatement/PRISMAStatement> [accessed 2022-07-11]
22. PRISMA checklist. PRISMA - Transparent Reporting of Systematic Reviews and Meta-Analyses. 2020. URL: <http://www.prisma-statement.org/PRISMAStatement/Checklist> [accessed 2022-07-12]
23. World Health Organization. Regional Office for Africa. Guide for the establishment of health observatories. World Health Organization. Regional Office for Africa. 2016. URL: <https://apps.who.int/iris/handle/10665/246123> [accessed 2023-04-14]
24. Equator Network homepage. Equator Network. URL: <https://www.equator-network.org/> [accessed 2022-07-19]
25. Brand RA. Standards of reporting: the CONSORT, QUORUM, and STROBE guidelines. *Clin Orthop Relat Res* 2009 Jun 19;467(6):1393-1394 [[FREE Full text](#)] [doi: [10.1007/s11999-009-0786-x](https://doi.org/10.1007/s11999-009-0786-x)] [Medline: [19296187](https://pubmed.ncbi.nlm.nih.gov/19296187/)]
26. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)* 2015 Mar 23;3(1):1052 [[FREE Full text](#)] [doi: [10.13063/2327-9214.1052](https://doi.org/10.13063/2327-9214.1052)] [Medline: [25992385](https://pubmed.ncbi.nlm.nih.gov/25992385/)]
27. Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: an overview. *Indian J Pharmacol* 2012 Mar;44(2):168-172 [[FREE Full text](#)] [doi: [10.4103/0253-7613.93842](https://doi.org/10.4103/0253-7613.93842)] [Medline: [22529469](https://pubmed.ncbi.nlm.nih.gov/22529469/)]
28. Berger A, Rustemeier AK, Göbel J, Kadioglu D, Britz V, Schubert K, et al. How to design a registry for undiagnosed patients in the framework of rare disease diagnosis: suggestions on software, data set and coding system. *Orphanet J Rare Dis* 2021 May 01;16(1):198 [[FREE Full text](#)] [doi: [10.1186/s13023-021-01831-3](https://doi.org/10.1186/s13023-021-01831-3)] [Medline: [33933089](https://pubmed.ncbi.nlm.nih.gov/33933089/)]
29. Mukhina AA, Kuzmenko NB, Rodina YA, Kondratenko IV, Bologov AA, Latysheva TV, et al. Primary immunodeficiencies in Russia: data from the national registry. *Front Immunol* 2020 Aug 6;11:1491 [[FREE Full text](#)] [doi: [10.3389/fimmu.2020.01491](https://doi.org/10.3389/fimmu.2020.01491)] [Medline: [32849507](https://pubmed.ncbi.nlm.nih.gov/32849507/)]
30. Licht C, Ardissino G, Ariceta G, Cohen D, Cole JA, Gasteyger C, et al. The global aHUS registry: methodology and initial patient characteristics. *BMC Nephrol* 2015 Dec 10;16(1):207 [[FREE Full text](#)] [doi: [10.1186/s12882-015-0195-1](https://doi.org/10.1186/s12882-015-0195-1)] [Medline: [26654630](https://pubmed.ncbi.nlm.nih.gov/26654630/)]
31. Messiaen C, Le Mignot L, Rath A, Richard JB, Dufour E, Ben Said M, et al. CEMARA: a web dynamic application within a N-tier architecture for rare diseases. *Stud Health Technol Inform* 2008;136:51-56 [Medline: [18487707](https://pubmed.ncbi.nlm.nih.gov/18487707/)]
32. Messiaen C, Racine C, Khatim A, Soussand L, Odent S, Lacombe D, AnDDI-Rares network, et al. 10 years of CEMARA database in the AnDDI-Rares network: a unique resource facilitating research and

- epidemiology in developmental disorders in France. *Orphanet J Rare Dis* 2021 Aug 04;16(1):345 [[FREE Full text](#)] [doi: [10.1186/s13023-021-01957-4](https://doi.org/10.1186/s13023-021-01957-4)] [Medline: [34348744](#)]
33. Stanimirovic D, Murko E, Battelino T, Groselj U. Development of a pilot rare disease registry: a focus group study of initial steps towards the establishment of a rare disease ecosystem in Slovenia. *Orphanet J Rare Dis* 2019 Jul 09;14(1):172 [[FREE Full text](#)] [doi: [10.1186/s13023-019-1146-x](https://doi.org/10.1186/s13023-019-1146-x)] [Medline: [31288838](#)]
34. Pastores GM, Arn P, Beck M, Clarke JT, Guffon N, Kaplan P, et al. The MPS I registry: design, methodology, and early findings of a global disease registry for monitoring patients with Mucopolysaccharidosis Type I. *Mol Genet Metab* 2007 May;91(1):37-47 [doi: [10.1016/j.ymgme.2007.01.011](https://doi.org/10.1016/j.ymgme.2007.01.011)] [Medline: [17336562](#)]
35. Stirnadel-Farrant H, Kudari M, Garman N, Imrie J, Chopra B, Giannelli S, et al. Gene therapy in rare diseases: the benefits and challenges of developing a patient-centric registry for Strimvelis in ADA-SCID. *Orphanet J Rare Dis* 2018 Apr 06;13(1):49 [[FREE Full text](#)] [doi: [10.1186/s13023-018-0791-9](https://doi.org/10.1186/s13023-018-0791-9)] [Medline: [29625577](#)]
36. Tingley K, Lamoureux M, Pugliese M, Geraghty MT, Kronick JB, Potter BK, Canadian Inherited Metabolic Diseases Research Network. Evaluation of the quality of clinical data collection for a pan-Canadian cohort of children affected by inherited metabolic diseases: lessons learned from the Canadian Inherited Metabolic Diseases Research Network. *Orphanet J Rare Dis* 2020 Apr 10;15(1):89 [[FREE Full text](#)] [doi: [10.1186/s13023-020-01358-z](https://doi.org/10.1186/s13023-020-01358-z)] [Medline: [32276663](#)]
37. Subirats L, Conesa J, Armayones M. Biomedical holistic ontology for people with rare diseases. *Int J Environ Res Public Health* 2020 Aug 19;17(17):6038 [[FREE Full text](#)] [doi: [10.3390/ijerph17176038](https://doi.org/10.3390/ijerph17176038)] [Medline: [32825147](#)]
38. Kalankesh LR, Dastgiri S, Rafeey M, Rasouli N, Vahedi L. Minimum data set for cystic fibrosis registry: a case study in Iran. *Acta Inform Med* 2015 Feb;23(1):18-21 [[FREE Full text](#)] [doi: [10.5455/aim.2015.23.18-21](https://doi.org/10.5455/aim.2015.23.18-21)] [Medline: [25870486](#)]
39. Shahmoradi L, Ehtesham H, Mehraeen E, Rostampour N, Tahmasbian S, Ghasempour M. Instructional design, development and evaluation of congenital hypothyroidism registry system. *Webology* 2019 Dec 30;16(2):275-288 [doi: [10.14704/web/v16i2/a203](https://doi.org/10.14704/web/v16i2/a203)]
40. McCann LJ, Arnold K, Pilkington CA, Huber AM, Ravelli A, Beard L, et al. Developing a provisional, international minimal dataset for Juvenile Dermatomyositis: for use in clinical practice to inform research. *Pediatr Rheumatol Online J* 2014 Jul 21;12(1):31 [[FREE Full text](#)] [doi: [10.1186/1546-0096-12-31](https://doi.org/10.1186/1546-0096-12-31)] [Medline: [25075205](#)]
41. McCann LJ, Kirkham JJ, Wedderburn LR, Pilkington C, Huber AM, Ravelli A, et al. Development of an internationally agreed minimal dataset for juvenile dermatomyositis (JDM) for clinical and research use. *Trials* 2015 Jun 12;16(1):268 [[FREE Full text](#)] [doi: [10.1186/s13063-015-0784-0](https://doi.org/10.1186/s13063-015-0784-0)] [Medline: [26063230](#)]
42. Huemer M, Diodato D, Martinelli D, Olivieri G, Blom H, Gleich F, EHOD consortium, et al. Phenotype, treatment practice and outcome in the cobalamin-dependent remethylation disorders and MTHFR deficiency: data from the E-HOD registry. *J Inherit Metab Dis* 2019 Mar 17;42(2):333-352 [doi: [10.1002/jimd.12041](https://doi.org/10.1002/jimd.12041)] [Medline: [30773687](#)]
43. Derayeh S, Kazemi A, Rabiei R, Hosseini A, Moghaddasi H. National information system for rare diseases with an approach to data architecture: a systematic review. *Intractable Rare Dis Res* 2018 Aug;7(3):156-163 [[FREE Full text](#)] [doi: [10.5582/irdr.2018.01065](https://doi.org/10.5582/irdr.2018.01065)] [Medline: [30181934](#)]

44. Taruscio D, Vittozzi L, Choquet R, Heimdal K, Iskrov G, Kodra Y, et al. National registries of rare diseases in Europe: an overview of the current situation and experiences. *Public Health Genomics* 2015 Sep 9;18(1):20-25 [doi: [10.1159/000365897](https://doi.org/10.1159/000365897)] [Medline: [25228300](https://pubmed.ncbi.nlm.nih.gov/25228300/)]
45. Opladen T, Gleich F, Kozich V, Scarpa M, Martinelli D, Schaefer F, et al. U-IMD: the first Unified European registry for inherited metabolic diseases. *Orphanet J Rare Dis* 2021 Feb 18;16(1):95 [FREE Full text] [doi: [10.1186/s13023-021-01726-3](https://doi.org/10.1186/s13023-021-01726-3)] [Medline: [33602304](https://pubmed.ncbi.nlm.nih.gov/33602304/)]
46. Meissner Y, Fischer-Betz R, Zink A, Strangfeld A. Response to: 'correspondence on 'EULAR recommendations for a core data set for pregnancy registries in rheumatology' by De Cock. *Ann Rheum Dis* 2023 Feb 25;82(2):e45 [doi: [10.1136/annrheumdis-2020-219478](https://doi.org/10.1136/annrheumdis-2020-219478)] [Medline: [33239275](https://pubmed.ncbi.nlm.nih.gov/33239275/)]
47. Ulin PR, Robinson ET, Tolley EE. *Qualitative Methods in Public Health A Field Guide for Applied Research*. Hoboken, New Jersey, United States: Wiley; 2004.
48. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat* 2012 May 06;33(5):803-808 [doi: [10.1002/humu.22078](https://doi.org/10.1002/humu.22078)] [Medline: [22422702](https://pubmed.ncbi.nlm.nih.gov/22422702/)]
49. Orphanet nomenclature for coding and associated tools. Orphanet. URL: <https://www.orphadata.com/orphanet-nomenclature-for-coding/> [accessed 2022-07-21]
50. NIH U.S. National Library of Medicine. 2005 Sep 20. URL: <https://clinicaltrials.gov/ct2/show/NCT00196742/> [accessed 2023-04-24]
51. Orth M, European Huntington's Disease Network, Handley OJ, Schwenke C, Dunnett S, Wild EJ, et al. Observing Huntington's disease: the European Huntington's Disease Network's REGISTRY. *J Neurol Neurosurg Psychiatry* 2011 Dec 19;82(12):1409-1412 [doi: [10.1136/jnnp.2010.209668](https://doi.org/10.1136/jnnp.2010.209668)] [Medline: [21097549](https://pubmed.ncbi.nlm.nih.gov/21097549/)]
52. Kodra Y, Minelli G, Rocchetti A, Manno V, Carinci A, Conti S, National Rare Diseases Registry Collaborating Group. The Italian National Rare Diseases Registry: a model of comparison and integration with Hospital Discharge Data. *J Public Health (Oxf)* 2019 Mar 01;41(1):46-54 [doi: [10.1093/pubmed/fox176](https://doi.org/10.1093/pubmed/fox176)] [Medline: [29294017](https://pubmed.ncbi.nlm.nih.gov/29294017/)]
53. Ambrosini A, Calabrese D, Avato FM, Catania F, Cavaletti G, Pera MC, et al. The Italian neuromuscular registry: a coordinated platform where patient organizations and clinicians collaborate for data collection and multiple usage. *Orphanet J Rare Dis* 2018 Oct 04;13(1):176 [FREE Full text] [doi: [10.1186/s13023-018-0918-z](https://doi.org/10.1186/s13023-018-0918-z)] [Medline: [30286784](https://pubmed.ncbi.nlm.nih.gov/30286784/)]
54. Ruiz E, Ramalle-Gómara E, Quiñones C, Spain RDR Working Group. Record linkage between hospital discharges and mortality registries for motor neuron disease case ascertainment for the Spanish National Rare Diseases Registry. *Amyotroph Lateral Scler Frontotemporal Degener* 2014 Jun 18;15(3-4):275-278 [doi: [10.3109/21678421.2014.890226](https://doi.org/10.3109/21678421.2014.890226)] [Medline: [24641576](https://pubmed.ncbi.nlm.nih.gov/24641576/)]
55. Rubinstein YR, McInnes P. NIH/NCATS/GRDR® common data elements: a leading force for standardized data collection. *Contemp Clin Trials* 2015 May;42:78-80 [FREE Full text] [doi: [10.1016/j.cct.2015.03.003](https://doi.org/10.1016/j.cct.2015.03.003)] [Medline: [25797358](https://pubmed.ncbi.nlm.nih.gov/25797358/)]
56. Grinspan ZM, Patel AD, Shellhaas RA, Berg AT, Axeen ET, Bolton J, Pediatric Epilepsy Learning Healthcare System. Design and implementation of electronic health record common data elements for pediatric epilepsy: foundations for a learning health care system. *Epilepsia* 2021 Jan 24;62(1):198-216 [FREE Full text] [doi: [10.1111/epi.16733](https://doi.org/10.1111/epi.16733)] [Medline: [33368200](https://pubmed.ncbi.nlm.nih.gov/33368200/)]



57. Murphy MS, Fell DB, Sprague AE, Corsi DJ, Dougan S, Dunn SI, et al. Data resource profile: better outcomes registry and network (BORN) Ontario. *Int J Epidemiol* 2021 Nov 10;50(5):1416-147h [[FREE Full text](#)] [doi: [10.1093/ije/dyab033](https://doi.org/10.1093/ije/dyab033)] [Medline: [34097034](https://pubmed.ncbi.nlm.nih.gov/34097034/)]
58. Mowry EM, Bermel RA, Williams JR, Benzinger TL, de Moor C, Fisher E, et al. Harnessing real-world data to inform decision-making: multiple sclerosis partners advancing technology and health solutions (MS PATHS). *Front Neurol* 2020 Aug 7;11:632 [[FREE Full text](#)] [doi: [10.3389/fneur.2020.00632](https://doi.org/10.3389/fneur.2020.00632)] [Medline: [32849170](https://pubmed.ncbi.nlm.nih.gov/32849170/)]
59. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021 Mar 29;372:n160 [[FREE Full text](#)] [doi: [10.1136/bmj.n160](https://doi.org/10.1136/bmj.n160)] [Medline: [33781993](https://pubmed.ncbi.nlm.nih.gov/33781993/)]
60. Vasant D, Chanas L, Hanauer M, Malone J. ORDO: an ontology connecting rare disease, epidemiology and genetic data. In: *Proceedings of the Bio-Ontology @ ISMB 2014*. 2014 Presented at: Bio-Ontology @ ISMB 2014; Jul 13-15, 2014; Boston MA USA

## 6 MAPPING, INFRASTRUCTURE, AND DATA ANALYSIS FOR THE BRAZILIAN NETWORK OF RARE DISEASES: PROTOCOL FOR THE RARASNET OBSERVATIONAL COHORT STUDY

### *6.1 Introduction*

#### **6.1.1 Background**

A rare disease (RD) is a medical condition with low prevalence compared to diseases prevalent in the general population, but there is no consensus on its definition [1]. The European Union describes a RD as a disease that affects no more than 1 person in 2000 [2]. In the United States, the Rare Disease Act of 2002 considers a RD any condition that affects less than 200,000 people across the country, or 1 person in 1500 [3]. In Japan, a RD is considered to affect less than 50,000 people (ie, 1 in 2500 people) [4]. On the other hand, in Latin America, there is no consensus on the definition of RDs in terms of numbers. Each country has its own definition according to its public policies, decrees, the existence of adequate treatments, or the severity of the disease [5].

Although individually rare, RDs collectively affect up to 10% of the population. Thus, rare diseases have a significant impact on health, and health professionals must be familiar with their diagnosis, management, and treatment [6]. It is estimated that there are up to 8000 rare diseases identified, 80% of which are of genetic origin [7]. In Brazil, the Ministry of Health defines a rare disease using the criteria of the World Health Organization (WHO), that is, 1.3 cases per 2000 individuals [8]. In 2014, the Brazilian Policy of Comprehensive Care for People with Rare Diseases was established within the scope of the Unified Health System (Sistema Único de Saúde [SUS]) [9].

To overcome this informational barrier, the WHO recommends that research involving the study of the process dynamics of RDs at the national level be financed by public agencies. The formation of large multidisciplinary networks is part of a fundamental process to encourage the collaboration of medical specialists, referral centers, and patient groups [10]. Providing an infrastructure for new mechanisms promoting the translation of basic research into clinically important products is still a priority. One of the most important opportunities addressed by the WHO reinforces support for “networks of excellence that focus on research infrastructures; research, infrastructure, and implementation of guidelines for medical and psychosocial care; [and] methods to provide easy access to health care available to patients, regardless of where they live” [11].

The first full version of a report relating to the concept of a health regionalized network emerged after World War I due to the consequent need for changes in the social protection system, presenting better ways for health services organization. Almost one century later, the original purpose remains very similar, as to establish a uniform system of clinical histories was declared a crucial reason for the better integration of different health service levels [12].

The proposal to integrate health networks gained momentum with the advent of health observatories. The emergence of these epidemiological monitoring centers is related to rapid changes in the health sector, including the need to monitor and assess the impact of health programs and public health policies, the advent of informational intelligence, digital health, and health knowledge management. The health observatories' functions include locating, gathering, analyzing, synthesizing, and disseminating data on the health status of a population, in addition to establishing partnerships and contacts with other agencies involved in the health of that region [13].

Independent of which model is used, the majority of countries today have digital systems to manage their data according to their health network structures. According to WHO recommendations, although digital health interventions are not enough on their own, when combined with health professionals, they are vital tools to promote health quality [11]. While data are still the main asset in the current digital world, many institutions do not yet fully understand the need and advantages of sharing their data with other organizations [14].

Cross-institutional sharing of health data is a challenge because many institutions are unwilling to share data due to privacy concerns or the fear of giving other institutions competitive advantages and, at an operational level, because of mistrust of technical barriers (there is no common platform for sharing these complex and heterogeneous data). On the other hand, overcoming this challenge may lead to better clinical effectiveness and improved clinical research [15]. Even if these concerns can be beaten, there is no consensus about the exact technical infrastructure needed to support such an effort.

Studies have shown that there remain a set of meaningful hurdles to achieving the desired benefits of health care data exchange. For example, failing to secure the patient record has financial and legal consequences as well as the negative potential to impact patient care. Thus, possible repercussions of a breach are a discouragement to swapping data. To avoid ethical and legal consequences for institutions, anonymization and privacy must be ensured for sensitive data, making them available only to authorized persons. Further, data anonymity could help improve the research area by removing identifiable information and sharing only limited data [15,16].

Another significant barrier that health networks need to face is adequate technological infrastructure. In addition to the scarce and fragmented availability of data on RDs [17], several technical aspects are common, like a centralized data source, which represents a high-security risk due to the susceptibility for malicious attacks to a nonredundant authority. A secure channel to send data to other organizations is another feature that institutions must address to avoid unauthorized access [14].

Due to the complexity of data in the health domain, achieving full interoperability is a hard task. Heterogeneous structures and data diversity decrease the accuracy of analysis and reduce understanding of information. To face this issue, several entities have created standards for data exchange. However, there is no consensus on the most adequate ones [18]. In Brazil, the Ministry of Health Ordinance 2073 of August 31, 2011, regulates the use of interoperability and health information standards for health information systems within the scope of the SUS; at the municipal, state, and federal levels; and in private systems and the private health sector [19].

In this sense, for there to be semantic interoperability between independent systems, it is necessary to standardize two aspects of information: the structure of the information and the semantic representation of the information. The information structure concerns the information and knowledge models that allow the systems to exchange data, formed in larger structures such as documents, correctly. The semantic representation of information includes terminologies, ontologies, and controlled vocabularies [20].

This framework can mitigate the several barriers preventing data access for research support agencies, academics, managers, and health professionals, such as the noncomputerization of processes, heterogeneity and duplicity of data in health information systems, and existence of a large amount of isolated data in databases accessible only in a certain context, usually to answer specific questions in particular research [17].

These factors often cause problems in the quality of information, making it difficult to coordinate and evaluate data in a research network linked to a rare disease patient care network, so it is not possible to use the data to assist in the decision-making process. Therefore, in health care, decision support tools are essential to guide the practice of health care and support the decisions of managers who will directly influence the quality of care provided to patients with RDs [21].

This paper presents a subproject belonging to the main project, entitled the Brazilian Rare Disease Network, funded by the Brazilian Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico [CNPq]), with

a 2-year forecast. The main project is a mixed prospective and retrospective observational cohort study to map the landscape of rare diseases in Brazil. RARASnet is responsible for infrastructure and data analysis to provide health indicators and support the construction, organization, and monitoring of this patient network. Although the objectives are collaborative, different teams are responsible for specific parts.

### **6.1.2 Related Work**

RDs represent a major challenge for the organization of health care. The cooperation of health service professionals, civil society, and academia is essential to overcome this challenge. In recent years, different collaborative networking initiatives have emerged. Collaborative networks are of great value for science and technology institutions to share, generate, and disseminate new knowledge that can lead to innovations and form a solid basis for a national care network [22-24]. In this scenario, records represent an important tool in acquiring the necessary knowledge about the clinical form and natural history of patients with RD.

Maintaining records of epidemiological data also contributes to the planning of public health programs, which in some cases requires supranational coordination. Thus, European Reference Networks (ERNs) were organized, supported by a series of rules and guidelines that provide a cohesive structure for sharing good practices of diagnosis, treatment, and standardization of recommended approaches for RDs. The promotion of ERNs contributed to the identification of already established centers of expertise and encouraged the voluntary participation of health service professionals in ERNs dedicated to specific groups of RDs [22].

Italy was one of the first European countries to develop specific RD regulations. The success of the experience in Italy is exemplified by the regional network of Piedmont and Valle d'Aosta, where networked activities have provided several benefits, such as improvement of multidisciplinary knowledge, provision of quality care, and reduced cost of therapeutic mobility [22,23]. To produce epidemiological evidence on RDs and support health service policy and planning, Italy also assessed the integrity and consistency of procedures carried out from its national registry and found that the data quality still represents a limitation to any solid epidemiological estimate [22].

After comparing the outcome of patients with primary systemic amyloidosis in a referral center with the population of this same Italian network, another study in the country found that the patients observed by the network had a diagnosis 4 months earlier than those seen in the

reference center. In addition to the rapid dissemination of knowledge pointed out as the main cause of this difference, important epidemiological differences were observed, which further reinforces the need for the standardization of reliable prognosis and the administration of clinical trial results [22,24].

According to the French National Plan for Rare Diseases, the first step in identifying patients with RD who are eligible for clinical trials or cohort studies is the definition of a minimum set of national data. In addition, providing reference centers with information technology (IT) tools contributes to the improvement of the service and research of RDs. Thus, according to international regulations on privacy and intellectual property and based on interoperability and semantics standards, the construction of the French model allowed data sharing in a national network composed of 131 centers specialized in RDs [22,24-26].

One of these French national reference centers went further and created a web-based medical archive of pediatric interstitial lung diseases. The construction of a national database made it possible to centralize and serve various stakeholders, such as researchers, clinicians, epidemiologists, and the pharmaceutical industry. Consequently, with the increasing engagement of new participants and the creation of committees to control data quality, there was an increase in the accuracy of the information provided, and several alternative solutions, depending on local possibilities, were configured [27]. Similar initiatives have also taken place in Germany [28] and the United Kingdom [29].

To not only collect records but also analyze them, a conceptual and digital framework based on the Asia-Pacific Economic Cooperation Rare Disease Action Plan has been articulated [26]. A proposal for a rare disease registry and analytical platform aims to assist in clinical decision making and improve the design and delivery of health services [30].

On the other side of the ocean, the National Center for the Advancement of Translational Sciences, one of the 27 departments of the US National Institutes of Health (NIH), maintains initiatives that aim to enhance the research of rare diseases, such as the promotion of information sharing and the construction of multidisciplinary collaborations. The Rare Diseases Clinical Research Network, for example, despite being formed by distinct clinical research consortia, shares the same data coordination and management center [31].

This management is only possible due to the availability of a genomic database maintained by the Genetic and Rare Diseases Information Center and the RD record program based on international standards and sharing (eg, Health Level Seven, Human Phenotype Ontology [HPO]) as well as the toolkit for the development of patient-focused therapies (National Center for Advancing Translational Sciences Toolkit), represented by an information

portal with guidelines for the development process of research and partnerships with the NIH and the Food and Drug Administration [32].

Worldwide, these two actions integrated not only clinical and epidemiological data and records but also information from biorepositories of biological samples for rare biospecimens (RD-HUB) [33], and they created an integrated platform that connects databases, records, biobanks, and bioinformatics clinics for rare disease research (RD-Connect) [34]. To address the quality of these data, several models and tools have been developed worldwide [34,35]. An assessment approach for diagnosing rare diseases based on Unified Modeling Language and ontologies, called FindZebra, improves the quality of diagnosis compared with standard search tools [34-36].

To become more than a search engine, decision support systems for clinical diagnosis have incorporated artificial intelligence and natural language processing techniques to provide more accurate and useful systems [37]. By having the infrastructure established according to the Ministry of Health, we can focus on the data analysis. Data science has been playing a major role in retrieving insights from patient reports and human and technical resources. Thus, the main goals of this project are described in the following section.

### **6.1.3 Objectives**

The primary objective of this study is to identify the essential elements for mapping, infrastructure, and data analysis for the Brazilian Network of Rare Diseases. Secondary objectives are to (1) create and implement a system that allows the integration of data available in different systems of health care, social assistance, and epidemiology of RD cases (a shared electronic medical record), simplifying the access to patient data via the web by health care stakeholders; (2) promote interoperability between health information systems through the use of the Semantic Web combined with traditional communication and data exchange techniques for functional and semantic interoperability and the integration of databases to improve the management of health services data; (3) develop a single and complete database using cloud computing with an access hierarchy and well-defined security rules by building a ubiquitous platform capable of providing access services and adding syntactic and semantic value to data, covering innovative techniques such as the use of blockchain for cloud computing; and (4) develop an evidence-based portal with national protocols for monitoring and analyzing data collected or produced in several RD patient care settings, incorporating data processing,

analysis, and machine learning techniques to assess the clinical situation and possible patient risks in real time.

## **6.2 Methods**

### **6.2.1 Brazilian Rare Disease Network**

Typically, information technology investigations can be distinguished as applied basic research. Basic research is scientific research focused on improving the understanding of phenomena and events [38]. Applied research uses scientific studies to develop technologies and methods to intervene in natural or other phenomena, aiming to improve human interaction with such phenomena [39].

As mentioned, the study described is part of a larger project, entitled the Brazilian Rare Disease Network, with a collection of quantitative data coupled with an innovation proposal, the creation of an epidemiological surveillance service network involving university hospitals, Reference Services for Neonatal Screening (Serviços de Referência em Triagem Neonatal [SRTNs]), and Reference Services for Rare Diseases (Serviços de Referência em Doenças Raras [SRDRs]) throughout the Brazilian territory.

Considering the goal of consolidating a national network of rare diseases that covers all regions of Brazil, this study has the participation of SRDRs, university hospitals that may or may not belong to the Brazilian Hospital Services Company (Empresa Brasileira de Serviços Hospitalares) network, and SRTNs. These centers are essential for building a national database that efficiently maps and represents the situation of the field of rare diseases in a country [40]. Brazil is divided into 5 regions (north, northeast, midwest, southeast, and south). The chosen participating centers are distributed across all Brazilian regions and are units of reference in health care for the population of their respective localities, according to the National Policy on Comprehensive Care of People with Rare Diseases [9].

Participating health centers are divided as follows by country regions: 6 centers in the north, 11 in the northeast, 6 in the midwest, 12 in the southeast, and 5 in the south. These include 16 Brazilian capitals that together have a total of 47 million people. In addition, as they are referral centers, they have the infrastructure to receive patients from smaller municipalities for the diagnosis and care of their population.

The area of care for people with rare diseases is structured into primary care and specialized care, following the Health Care Network (Rede de Atenção à Saúde) and the



Guidelines for the Comprehensive Care for People with Rare Diseases plan of the SUS. SRDRs are responsible for preventive, diagnostic, and therapeutic actions for individuals with rare diseases or at risk of developing them, according to care axes. The SRDSs have a network of Specialized Rehabilitation Centers (Centros Especializados em Reabilitação [CERs]), which can receive patients referred from SRDSs and assist in the rehabilitation of these patients [41].

The CERs are structural components of the National Policy on Comprehensive Care of People with Rare Diseases. According to the integrality of care, these centers perform treatment, concession, adaptation, and maintenance of assistive technology, constituting a reference for the health care network in the territory [42]. SRDRs and CERs work together with university hospitals to promote comprehensive and universal care for rare disease patients.

The traditional concept defines a university hospital as an institution that is characterized by four traits: being an extension of a health teaching establishment (of a medical school, for example), providing university training in the health field, being officially recognized as a teaching hospital and subject to the supervision of competent authorities, and providing more complex medical care (tertiary level) to a portion of the population and being able to receive patients from SRTNs [42,43].

The SRTNs are units with multiprofessional health teams accredited and specialized in assistance, follow-up, treatment, and redirection of newborn patients diagnosed with pathologies such as phenylketonuria, congenital hypothyroidism, sickle cell diseases, biotinidase deficiency, congenital adrenal hyperplasia, and cystic fibrosis. Such pathologies are detected in the SRTN's own or an outsourced laboratory, according to the rules established in the National Neonatal Screening Program [44].

Initially, the 3 main collaborator groups consist of 17 university hospitals, 6 SRTNs, and 17 SRDRs. The effective consolidation of the Brazilian network of rare diseases, based on the mapping of these services, depends on 3 steps: (1) approval by the ethics committee of the coordinating institution of the project, (2) approval of the local ethics committees of each participating institution, and (3) consolidation of the human resources participating in each institution through the institutional consent form.

The first step has already been completed and the others are in progress. Any divergence in these steps results in the exclusion of the participating center from the project. While these steps are in progress, representatives of all participating institutions meet monthly—on the second Saturday of the month in the morning—to discuss and structure the other activities of the project. Additionally, institutions must disseminate and invite partner services to participate

in the initiative. The structuring and alignment of the final group of participants in the Brazilian network of rare diseases was finalized in August 2020.

### **6.2.2 Ethical Considerations**

The National Network of Rare Diseases project was approved (Edital No. 25/2019) from CNPq, with financial support from the Ministry of Health in the amount of R \$3.5 million (US \$662,139.10) [45]. Moreover, the main project was sent to the research ethics committee of Porto Alegre Clinical Hospital of the Federal University of Rio Grande do Sul (Hospital de Clínicas de Porto Alegre da Universidade Federal do Rio Grande do Sul) through Plataforma Brasil, a Brazilian platform of the Ministry of Health projects. The research ethics committee of Porto Alegre Clinical Hospital analyzed the research project (under code number 33970820.0.1001.5327 of Presentation Certificate for Ethical Appreciation). The research was approved (opinion number 4.225.579) on August 14, 2020.

To ensure the anonymity of patients while making it possible to track them if necessary, a password will be created for all patients, consisting of the first 2 letters of the city followed by the center number with 2 digits (from 01 for each city) and a 2-digit sequence for the patient's number. The rights, safety, and well-being of the subjects involved in the study will be the most important considerations and should prevail over the interests of science and society.

Considering the governmental efforts (ConecteSUS) [46,47], we similarly propose the use of a permissioned distributed blockchain solution that uses a key pair (private and public key) and a symmetrical consortium key for data encryption. A consortium distributed storage network will be established, consisting of research centers and other approved stakeholders throughout Brazil [48].

Authentication, authorization, integrity, and confidentiality verification mechanisms will be implemented through the establishment of a security layer. Thus, the security structure presented in this project aims to protect sensitive data for interoperability purposes. All computational techniques that support the solution, such as encryption and hashing, are well-known technologies that, when combined, can offer robust security features. In this way, each candidate system to interoperate with the rare disease ecosystem can easily meet all the necessary technical requirements.

All data collection processes will match the novel Brazilian General Law of Data Protection (federal law No. 13.709/18) [49]. The law refers to the respect to user privacy,

transparency in the data collection, security, and prevention of damage in personal data. Since August 16, 2020, the law covers all national territory, and its violation can cause a warning, penalties, a data block, and suspension of the project [50]. As mentioned, the project will ensure the anonymity of the data during analysis. In addition, the IT team will present to all members of the network the definition and main aspects of the General Law on Protection of Personal Data (*Lei Geral de Proteção de Dados Pessoais*) using supporting materials and a patient consent form for data collection and usage, with full transparency.

### **6.2.3 RARASnet Project Management**

The project management will include the cooperation and execution of several activities, including technological and technical implementation, that must be harmonized. The technical IT group will coordinate activities related to electronic resources, such as data collection instruments design, database management, and data analysis. The IT team is also responsible for maintaining a communication channel with the project's principal investigators to receive clinical administrative and clinical research input.

A set of practices that merge development and operations (DevOps) will be used as a reference to standardize the development process and align activities of software engineering, infrastructure operation, and quality assurance. As an agile methodology, DevOps allows quick delivery of a small set of requirements from concept to deployment. The method also creates efficiency in results monitoring due to continuous integration and the appreciation of high-value feedback from all stakeholders [51].

For project management and to increase collaboration across team members, Trello (Atlassian) [52] will be used, which provides easy visualization of tasks and priorities, as well as a macrovision of development stages. The workflow of a data analysis project will follow the classic steps of a knowledge discovery in databases process [53], detailed in the following subsections.

### **6.2.4 Data Collection Procedures**

Initially, the instruments to be used in data collection will be framed, validated, and tested. These instruments should serve as a basis for the steps that involve the survey of retrospective data in the participating institutions and as a model for the stage involving the

prospective survey and analysis. Based on an initial report characterizing the informational maturity of the collaborating institutions, online training will be given to address the functioning of the data collection instrument developed, validated, and tested for the project's retrospective phase, and the same process will be carried out later in the prospective phase of the project.

The collection will be carried out through access to medical records, with data recording on portable computers acquired with funds from this proposal and carried out by fellows of the project with the support of researchers from each service. Data quality indicators will be monitored in this intervention, mainly about the difficulties encountered by institutions to codify the diseases in an interoperable way, ensuring the production of a reliable picture of the maturity of data collection of rare diseases in Brazil.

To ensure the monitoring of data quality indicators, an early hearing detection and intervention (EHDI) will be conducted, and dimensions such as completeness, uniqueness, timeliness, validity, accuracy, and consistency will be evaluated [54]. Elements not present in the EHDI, such as acceptability, reliability, and flexibility, will also be considered; the use of

dimensions will vary depending on the requirements of each center. These indicators were selected based on their importance in monitoring and evaluation in the National Policy on Comprehensive Care of People with Rare Diseases. It will also allow tracking results from the source to the national level and be indicative of data quality for all the indicators within a program area [55].

During data collection, phenotypic data will be described according to HPO terms, restricted to 5 terms per case, allowing the description of phenotypes of known syndromes. Information about the coding of the disease will also be presented, considering the name of the disease, the International Classification of Diseases 10th Revision (ICD-10), the Orpha number, and the gene name or symbol, thus allowing comparison with data from other platforms, such as Orphanet.

Data collection instruments (ie, case report forms [CRFs]) will be established by principal investigators and applied in distinct project phases, each with a specific objective. The development of all CRFs is guided by the National Policy of Comprehensive Care for People with Rare Diseases [56,57] in the context of the Brazilian Health Public System.

The main instruments are (1) a survey of the technical and technological resources of the participating research center, used to recognize needs and prepare and provide resources for data integration and collection across research centers; (2) a survey of procedures performed at participating centers, used to recognize the availability of technological resources for genetic diagnosis and human resources in the assistance of individuals with rare diseases; (3) a

retrospective collection of clinical data, that is, the characterization of the clinical profile of patients with RDs treated throughout the country in the last 2 years; and (4) a prospective collection of clinical data, that is, the follow-up of patients with the defined RD clinical profile treated throughout the country, for the identification of changes in the clinical profile, such as in diagnosis and treatment.

After the initial development, the validation phase will take place. Key researchers, along with main investigators, will perform several rounds of revision and validation for each CRF. This process will occur until researchers reach a consensus. Then, the final version of an instrument (usually a paper-based one) will be translated into an electronic-based version.

### **6.2.5 Computational Infrastructure and Data Collection Resources**

The study will rely on a computational infrastructure to satisfy technological needs during all project phases. First, cloud computing resources were acquired as an infrastructure as a service. This makes it possible to quickly scale up and down with demand. Additionally, the expense and complexity of buying, managing, and maintaining physical servers and other data center infrastructure are avoided [58].

In this case, the University of São Paulo provides a private cloud computing environment (interNuvem USP) and manages the whole infrastructure, while the project's owners only need to install, configure, and manage their own software, operating systems, and applications. Several resources, such as webs, database, and data collection servers, will be available to help deliver this project outcome.

During the project, it will be necessary to collect data using CRFs. To facilitate the creation of electronic CRFs and their distribution, REDCap (Research Electronic Data Capture) and KoBo Toolbox will be used as electronic data capture systems. REDCap was built in 2004 by a team at Vanderbilt University to enable classical and translational clinical research, basic science research, and general surveys, providing researchers with a tool for the design and development of electronic data capture tools [59].

KoBo Toolbox, developed by the Harvard Humanitarian Initiative, is a free and open-source suite of tools for field data collection and basic analysis. It was initially built for use in challenging environments in developing countries, but it can be extended to any type of research [60]. Both electronic data capture systems are free, although licensing is necessary for REDCap. After applying for a REDCap license of use, the RARAS REDCap Server was established,

which is now part of the REDCap Consortium, a community of experts and REDCap administrators [61]. KoBo Toolbox does not demand a licensing process and the software is publicly available for download and installation.

REDCap and KoBo Toolbox are integrated and can be used together. The first is used for data research, data storage, reporting, analysis, and management. The second is used exclusively in the data collection process as a front-end tool for final users, allowing responsive and offline data collection on any type of device without the need to install any third party or additional mobile app. After submitting a record in KoBo Toolbox, data are instantly synchronized with the REDCap database. This integration is possible due to a framework developed by the IT group.

### **6.2.6 Database Modeling**

By exploring the data sources of the Orphanet platform related to information on medicines and rare diseases, we started the modeling phase of the database. Additionally, materials were selected for the knowledge acquisition phase for the development of a computational ontology that will reuse the Orphanet Rare Disease Ontology (ORDO) [62], thus helping the classification and hierarchization of bibliographic data on the prevalence of these diseases. After this initial analysis to select the best attributes (variables) that represent this health domain and are aligned with the profiles of the participating centers, the second stage of modeling the database is expected to start [25].

The first step is important so that the system does not request variables that are not relevant to the study, reducing the time taken to collect patient data by the health professional. More specifically, conceptual modelers describe structure models in the form of entities, relationships, and constraints, as they can also describe behavioral or functional models in terms of states, transitions between states, and actions performed on states and transitions. Finally, they can describe interactions and user interfaces in terms of messages sent and received and information exchanged. At the end of the first stage, a system requirements document must be prepared, detailing all functional and nonfunctional aspects of the implementation and application layers [31].

To facilitate the understanding of the information flow and operational processes of the participating institutions, auxiliary diagrams will be produced using the Business Process Management Notation approach. Such documents will be used during the project to validate

the information from the services, which will also be useful for the implementation and maintenance phases of the database.

The second phase of the modeling, therefore, consists of mapping the model in the form of relational tables. To ensure data consistency, the mapping is done according to the rules of the relational model, which was chosen because of its simplicity and robustness and because it uses structured query language (SQL), which has become common in relational databases. To generate the first model of the proposed database, the MySQL Workbench (Oracle Corp) software will be used, which allows data management and SQL queries to be built and facilitates the administration, creation, and maintenance of several databases in the same location. In this way, the bank will be ready for use and its implementation will be dynamic, offering the scope for future updates and maintenance [32].

### **6.2.7 Data Quality Assurance**

As previously stated, both the retrospective and prospective phases will collect study data using the KoBo Toolbox electronic data capture tool and store them using the REDCap server hosted at the Ribeirão Preto Medical School, University of São Paulo, Brazil. The KoBo Toolbox online data entry system will minimize the data entry errors and facilitate the monitoring and quick resolution of queries and missing data.

The data collection tools will be reviewed by other researchers and pretested on a convenient sample of records and clinical settings. Reviewers will note their individual experience with both the definitional criteria and the time taken to collect and record data. Based on the final pretest, revisions will be made to both data collection instruments.

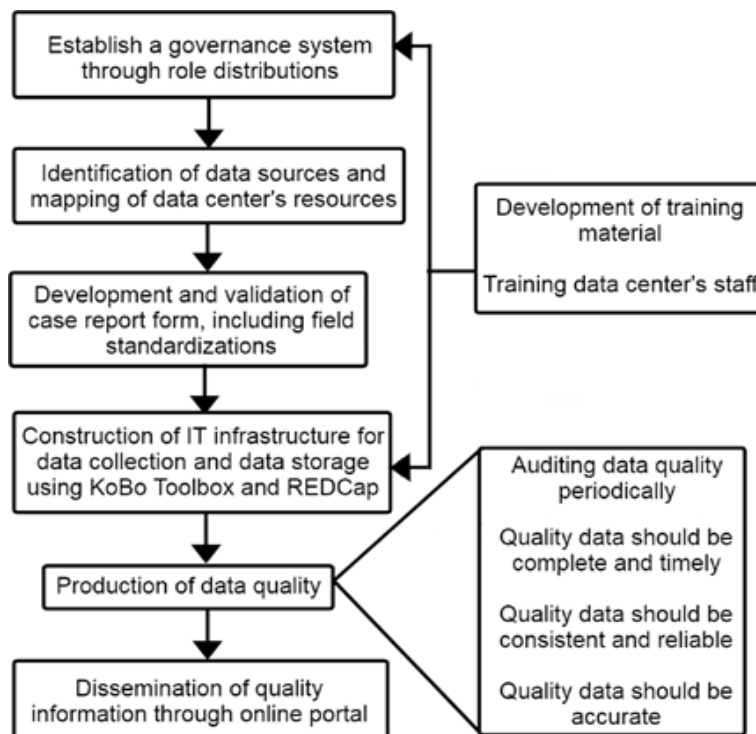
A manual of operations will be developed to minimize the need for judgment and interpretation by the data collectors and to increase the quality of data collection done by the health care center professionals. The manual of operations will include a description of the study in general terms, emphasize the importance of complete and accurate data, and foster the standardization of data collection.

The responsible health care center staff member will maintain a problem logbook to document unanticipated problems. Technical questions encountered in the field will be resolved through consultation with the technical team and researchers responsible for the project.

To ensure that the record quality fulfills all prerequisites described in the literature and the normative documents previously mentioned, we will follow a set of recommendations

described by the ERN in the RD-Connect framework, incorporating the indicators in each step of the process collection, storage, preprocessing, processing, and reporting [63,64]. Our plan will consider aspects such as governance standards, infrastructure in compliance with the FAIR principles (findable, accessible, interoperable, and reusable for humans and computers), didactic material, and informative documents, as well as personnel training and a data quality trail. The process and tools used for each level are presented in figure 7.

Figure 7 - Framework for quality management of rare disease registries. IT: information technology.



### 6.2.8 Data Management

Trained research nurses at the participating health facilities will use KoBo Toolbox data collection tools to collect data for both retrospective and prospective phases. All entries will be deidentified at the stage of data collection, and participants will be identifiable only by unique identification codes that are only accessible and known to the hospital coordinator. A customized data entry and monitoring system will be developed in the REDCap platform for this study. This data entry system will be password protected and accessible only to the database managers and study team. The system will be developed and coordinated by the study data management unit at the University of São Paulo, Brazil.



### 6.2.9 Portal Development and Data Analysis

After the identification and collection of the essential data, the IT team will be responsible for developing all the data analysis by the supervisors of specialists in the rare disease network. The analysis will serve as support for RARASnet specialists and patients to understand the main aspects of human and technical resources and the flow of rare diseases in Brazil. As retrospective and prospective data will be collected, they will serve as a base for the exploration of statistical and modeling computational methods, and with the validation of the results among specialists, the database will be incorporated into DATASUS and communicated in scientific reports and a web portal. This web portal will be one of the main practical contributions of this work. It refers to the Brazilian Digital Atlas of Rare Diseases, available through a health observatory, which aims to integrate structural information about the referential institutions working in rare diseases in the country and clinical information about the individuals assisted by these institutions. This building process will be done according to the guidelines proposed by WHO for the development of health observatories [65].

From that data organization, the analysis tools will be made available, providing health indicators to the managers (hospital, municipal, and regional). The main analyses of the web portal will be (1) the flow of patients, which will present the displacement of patients according to the place of origin and the hospital care through georeferenced maps and tables; (2) hospital indicators, which will provide the automated calculation of 31 hospital indicators, such as mortality, morbidity, capacity, and usability, aiming to observe and compare these indicators among institutions; (3) nosological profiles, which will highlight the hospital care of individuals, allowing for the characterization of morbidities in the rare disease community; (4) diseases sensitive to primary care, which will describe hospitalizations for morbidities related to primary care, facilitating the identification of hospitalization rates that could be avoided by strengthening primary care; (5) prediction of risk of death by the Charlson Comorbidity Index, which will provide the risk of death for patients according to their comorbidities; and (6) medical procedures, which will describe the surgical procedures performed, allowing the comparison between these procedures and the resources used [66-69].

The tools described will provide interactivity through consultation filters with spatial disaggregation (by region, health region, municipality, or a specific hospital) and temporal disaggregation. Thus, information will be able to be explored historically, geographically, and in real time, supporting different demands and decision making for rare diseases in Brazil. However, besides the web portal, which will contain general public information and resources,

we will also provide all the knowledge through videos and talks using didactic language to facilitate understanding.

### **6.3 Results**

Considering the objective list and proposed conceptual and technical model, RARASnet presents some outcomes of interest, both specific and collaborative, in 9 steps:

1. Survey epidemiology, clinical procedures, and therapeutic resources, such as the number of individuals with rare disease according to each diagnostic group, age, race, sex, and other features.
2. Create a national rare disease network with the participation of important university hospital health services in rare diseases to create a database of national rare diseases.
3. Cover all regions of Brazil concerning the main rare diseases, with institutes and number of cases stratified according to each type of condition.
4. Create a standard in sociodemographic, epidemiological, clinical, and therapeutics data with the advice of the specialists in the national rare disease network. The data should follow patterns proposed by the Ministry of Health guidelines and HPO terms.
5. Identify the type of treatments being applied in each center and those funded by SUS or by supplementary health. The goal is to have a quantitative analysis for each type of treatment to understand the overall status of rare disease in Brazil and perform public health policies.
6. Map existing diagnostic and technological resources within the network.
7. Map human resources, such as the quantity of workers and specialists available in the network in each region of Brazil.
8. Establish a network of partners to underpin collaborative studies concerning rare diseases.
9. Develop the online Brazilian Atlas of Rare Diseases according to the guidelines of the WHO [70] to help professionals, the general public, and political decisions.

The present project is in its initial stages, and a survey was completed by each reference center to evaluate the technical aspects of each health care center, such as the presence of computers, technical support staff, and a digitized system. Moreover, we are in the process of internally validating the collection instruments with specialists and principal investigators and preparing the pilot project to be carried out at the coordinating center for external validation. All the predicted methodological processes are shown in Figure 8.

For the participating centers that have already obtained the project approval from their respective ethics committees, we developed an initial data collection instrument to verify the technological infrastructure of each center and the way these institutions capture information related to rare diseases. This survey aims to list and categorize these institutions according to their methods of data storage and retrieval, which can be digital, through electronic medical records and management software, or analog, through paper-based record management.

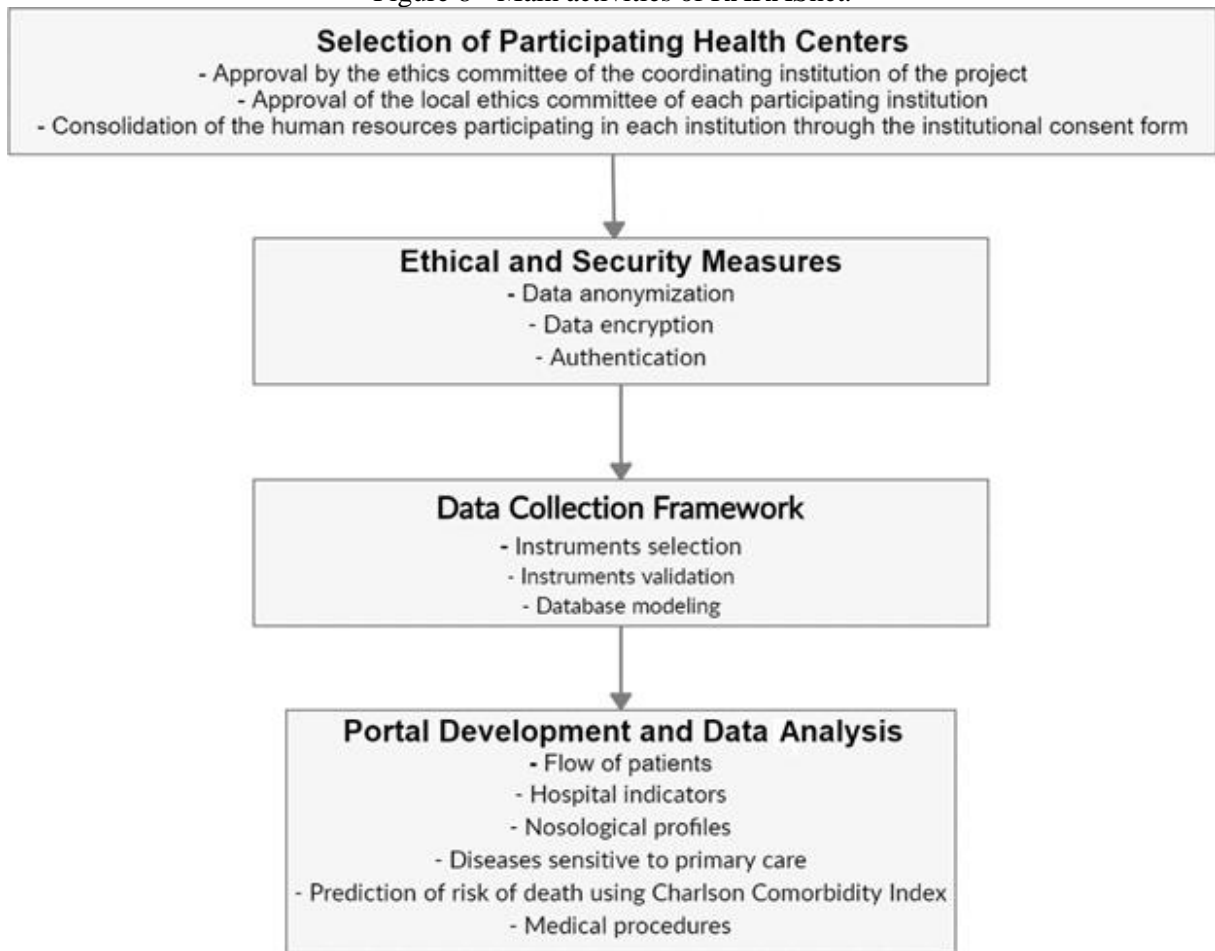
From a total of 40 participating centers so far, 39 have already responded to this initial survey, and among these, the results showed that 23 institutions (59%) have an electronic collection and recovery system and 16 institutions (41%) have a paper collection system. This initial survey is important for our IT team to plan the best clinical data collection approach for each institution during the project, aiming to minimize obstacles through an adequate and personalized collection proposal for each center.

One of the biggest challenges of this study is aligning data collection in all participating institutions so that the process of recovering data from medical records is standardized regardless of the storage support and the methods of retrieving information used in each health unit. In our scenario, based on 39 institutions, 14 (36%) health units extract information from medical records exclusively on paper, while 2 (5.1%) have a nonapplicable data recovery method; although they store their data in a system or on paper records, their recovery process does not fit into one of these methods (eg, using applications that are not for this purpose).

Later, we will standardize and analyze the clinical and epidemiological data and use these data to develop the national network for monitoring rare diseases, using the Digital Health Observatory to make the information available.

The project had its financing approved in December 2019. Retrospective data collection started in October 2020, and we expect to finish in January 2021. We will begin the prospective data collection in February 2021, and we expect to finish in June 2021. During the third quarter of 2020, we enrolled 40 health institutions from all regions of Brazil. We are currently receiving data to be analyzed. We expect the publication and dissemination of the findings in the second half of 2021.

Figure 8 - Main activities of RARASnet.



## 6.4 Discussion

This study is currently in its initial stage. We have performed a survey of technical aspects of health care centers (eg, support staff and technological infrastructure). A pilot data collection of clinical data carried out by specialists and principal investigators is planned to underpin the instruments' validation.

### 6.4.1 Main Problems Anticipated and Proposed Solutions

The heterogeneity of the data is intrinsically connected to the type of information generated by the health services, which is considered diverse and complex. Some of the main problems normally encountered when handling health data are the highly heterogeneous and sometimes ambiguous nature of medical language and its constant evolution; the huge amount of data generated constantly by the automation of hospital processes; the emergence of new

technologies; the need to process, analyze, and make decisions based on this information; and the need to ensure the safety of data related to patients [63].

To mitigate problems of heterogeneity and data standardization, we will make use of some semantic web technologies, which are presented as a fundamental approach to guarantee semantic interoperability and the integration of dispersed and isolated data sets. More specifically, we will use biomedical ontologies, which provide controlled vocabularies of scientific terminologies used to assist in the annotation of produced data, such as basic terms and their relations in a domain of interest, as well as rules to combine these terms and relations [64]. As mentioned, some of the ontologies used will be ORDO and HPO.

Once the ontologies are defined, it will be possible to perform the semantic markup on the collected records present in our relational database and provide a SPARQL Protocol and RDF Query Language access point to execute queries on the data set, allowing us to make available a set of data that can be extracted by different information systems, as long as they are connected to the web.

For security reasons related to the sensitivity of the stored information, direct access by external systems to the data structure is blocked by default. Therefore, an authorization layer will be built to support the authentication processes (validation of the identity of external systems). The authorization and protection of the information transmitted will use digital signature and hybrid encryption techniques, that is, a combination symmetric (unique key) and asymmetric (public and private key pair) encryption.

We believe that through these planned solutions, obtaining information from the set of data related to rare diseases in Brazil will become possible, allowing data to be shared, reused, analyzed, and applied in other information systems, either to improve the completeness of other bases or to produce relevant knowledge to support decision-making processes in the context of rare diseases.

#### **6.4.2 Applicability of the Results**

In the development of digital products and services for this project, all tools must ensure that users have the freedom to interactively navigate and filter data to visualize the analysis according to personal interests. For this, the Brazilian Digital Atlas of Rare Diseases will have a filter that allows spatial disaggregation (queries by regions, health regions, municipalities, or

a particular hospital) and temporal data. The filters will allow the user to set different visualization schemes without accessing the raw data and modifying the database.

It will also be possible to perform other types of data aggregation in queries, such as grouping by gender, age group, ICD-10, Orphacode, Online Mendelian Inheritance in Man (OMIM) [63], phenotypic characteristics, and other information that the health professionals involved consider relevant. It is important to emphasize that in the RD context, ICD-10 and OMIM are not able to cover all diseases with a unique identifier [71,72]. Thus, when ICD-10 is used as a filter, a further option box will be opened to distinguish between diseases with the same code. For OMIM, only genetic disorders are covered, and a note will be displayed on the website [63]. Orphacode [62], on the other hand, is the nomenclature that fits all RD diseases with a unique code due to its polyhierarchical nature [71].

This approach makes it possible to measure the performance of both the institution providing the health service and the care team. The analysis of efficiency and performance will be presented through dashboards and reports in real time, which can be used for the elaboration of new models based on the results.

The database of patients with rare diseases will allow an interactive epidemiological map and detail the care journey of the main rare diseases in Brazil. In this sense, it is expected that these developments can assist the evidence-based decision-making process for rare disease services in Brazil, bringing benefits to patients, health professionals, and managers.

#### **6.4.3 Plans for Validation, Dissemination, and Use of Project Results**

The dissemination of the results will include the production of scientific papers in periodicals relevant to the area and the realization of scientific dissemination to the direct target audience and collaborators through workshops and training to the participating centers. Aiming for project integration and sustainability, we will make the ontologies developed available in the international repository of biomedical ontologies, BioPortal. These artifacts will therefore be able to be used in other projects around the world and updated constantly. BioPortal is an open database that provides access to biomedical ontologies via web services, facilitating the participation of the scientific community in the evaluation and evolution of ontologies by suggesting additional resources for mapping terminologies and reviewing criteria and standards [73].

With the main results and interest topics, we intend to recruit a multidisciplinary panel for an e-Delphi [74] consensus-building exercise with the ad hoc team members. The e-Delphi method is an interactive structured communication technique to reach consensus on the responses, and it comprises an initial open round of questions to revise or suggest a list of potential items for scoring in the subsequent two scoring rounds.

Once results are validated, it is crucial “to design strategies and solutions to overcome bottlenecks that prevent proven and innovative public health interventions” from reaching the people who need them [75]. For this purpose, we intend to use the WHO toolkit for implementation research. One of the WHO toolkit topics describes how to plan a rigorous research project, including identifying implementation research outcomes, evaluating effectiveness, and making plans to scale up implementation in real-life settings [76].

Once we have the findings, we intend to analyze the implementation of these interventions and strategies. For this, the reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) framework [77] will be used to organize reviews of the existing literature on health promotion and disease management in different settings. RE-AIM is a tool used to translate research into action for digital technologies by measuring 5 essential dimensions for successful implementation: reach, effectiveness, adoption, implementation, and maintenance.

The overall goal of the RE-AIM framework is to encourage program planners, evaluators, readers of journal articles, donors, and policy makers to pay more attention to essential program elements, including external validity, which can improve the sustainable adoption and implementation of effective, generalizable, evidence-based interventions [78]. Finally, by applying the RE-AIM framework, we can emphasize responses to improve the chances that recommendations will have a positive and sustainable impact on public health.

## **6.5 References**

1. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, International Society for Pharmacoeconomics and Outcomes Research Rare Disease Special Interest Group. Rare Disease Terminology and Definitions-A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value Health* 2015 Sep;18(6):906-914 [FREE Full text] [doi: 10.1016/j.jval.2015.05.008] [Medline: 26409619]
2. Moliner A, Waligora J. The European Union Policy in the Field of Rare Diseases. *Adv Exp Med Biol* 2017;1031:561-587. [doi: 10.1007/978-3-319-67144-4\_30] [Medline: 29214592]

3. Rare Disease Act Of 2002, HR Rep No. 107-543. Congress.gov. 2002. URL: <https://www.congress.gov/congressional-report/107th-congress/house-report/543> [accessed 2021-01-08]
4. Singh J. The portal for rare diseases and orphan drugs. *J Pharmacol Pharmacother* 2013 Apr;4(2):168-169 [FREE Full text] [Medline: 23761721]
5. Mayrides M, Ruiz de Castilla EM, Szelepski S. A civil society view of rare disease public policy in six Latin American countries. *Orphanet J Rare Dis* 2020 Feb 27;15(1):60 [FREE Full text] [doi: 10.1186/s13023-020-1314-z] [Medline: 32106873]
6. Ng DM, Burnett JR, Bell DA, Hegele RA, Hooper AJ. Update on the diagnosis, treatment and management of rare genetic lipid disorders. *Pathology* 2019 Feb;51(2):193-201. [doi: 10.1016/j.pathol.2018.11.005]
7. Dharssi S, Wong-Rieger D, Harold M, Terry S. Review of 11 national policies for rare diseases in the context of key patient needs. *Orphanet J Rare Dis* 2017 Mar 31;12(1):63 [FREE Full text] [doi: 10.1186/s13023-017-0618-0] [Medline: 28359278]
8. Giugliani R, Vairo FP, Riegel M, de Souza CFM, Schwartz IVD, Pena SDJ. Rare disease landscape in Brazil: report of a successful experience in inborn errors of metabolism. *Orphanet J Rare Dis* 2016 Jun 10;11(1):76 [FREE Full text] [doi: 10.1186/s13023-016-0458-3] [Medline: 27282290]
9. Brasil. Portaria No 199, de 30 de janeiro de 2014. Ministério da Saúde. Brasília: Gabinete do Ministro; 2014 Jan 30. URL: [https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199\\_30\\_01\\_2014.html](https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html) [accessed 2020-09-15]
10. Priority diseases and reasons for inclusion. World Health Organization. URL: [https://www.who.int/medicines/areas/priority\\_medicines/Ch6\\_19Rare.pdf?ua=1](https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf?ua=1) [accessed 2020-09-07]
11. WHO Guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: <https://www.who.int/reproductivehealth/publications/digital-interventions-health-system-strengthening/en/> [accessed 2021-01-08]
12. Kuschnir R, Chorny AH. Redes de atenção à saúde: contextualizando o debate. *Ciência & Saúde Coletiva* 2010 Aug;15(5):2307-2316. [doi: 10.1590/s1413-81232010000500006]
13. Guide for the establishment of health observatories. World Health Organization Regional Office for Africa. 2016. URL: <https://apps.who.int/iris/handle/10665/246123> [accessed 2021-01-08]



14. Mamoshina P, Ojomoko L, Yanovich Y, Ostrovski A, Botezatu A, Prikhodko P, et al. Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare. *Oncotarget* 2018 Jan 19;9(5):5665-5690 [FREE Full text] [doi: 10.18632/oncotarget.22345] [Medline: 29464026]
15. Cichosz SL, Stausholm MN, Kronborg T, Vestergaard P, Hejlesen O. How to Use Blockchain for Diabetes Health Care Data and Access Management: An Operational Concept. *J Diabetes Sci Technol* 2019 Mar;13(2):248-253 [FREE Full text] [doi: 10.1177/1932296818790281] [Medline: 30047789]
16. Zhang A, Lin X. Towards Secure and Privacy-Preserving Data Sharing in e-Health Systems via Consortium Blockchain. *J Med Syst* 2018 Jun 28;42(8):140. [doi: 10.1007/s10916-018-0995-5] [Medline: 29956061]
17. de la Paz MP, Taruscio D, Groft SC. *Rare Diseases Epidemiology: Update and Overview*. Cham, Switzerland: Springer; 2017.
18. Liang X, Zhao J, Shetty S, Liu J, Li D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. 2017 Presented at: IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); Oct 8-13, 2017; Montreal, QC, Canada. [doi: 10.1109/pimrc.2017.8292361]
19. Braz RM, Oliveira PDTRD, Reis ATD, Machado NMDS. Avaliação da completude da variável raça/cor nos sistemas nacionais de informação em saúde para aferição da equidade étnico-racial em indicadores usados pelo Índice de Desempenho do Sistema Único de Saúde. *Saúde em Debate* 2013 Dec;37(99):554-562. [doi: 10.1590/s0103-11042013000400002]
20. Centro Nacional de Terminologias em Saúde: Planejamento Estratégico 2018-2021. Ministério de Saúde. 2018. URL: <https://portalarquivos2.saude.gov.br/images/pdf/2018/junho/14/planejamento-estrategico-centerms.pdf> [accessed 2021-01-08]
21. Lima VC, Lopes Rijo RPC, Pellison FC, de Lima RR, Cruz Correia RJ, Giuriati HT, et al. From guidelines to decision-making: using mobile applications and semantic web in the practical case of guides to support patients. *Procedia Computer Science* 2017;121:803-808. [doi: 10.1016/j.procs.2017.11.104]
22. Baldovino S, Moliner AM, Taruscio D, Daina E, Roccatello D. Rare Diseases in Europe: from a Wide to a Local Perspective. *Isr Med Assoc J* 2016 Jun;18(6):359-363 [FREE Full text] [Medline: 27468531]
23. Baldovino S, Menegatti E, Modena V, Maspoli M, Avanzi F, Roccatello D. Piedmont and Aosta Valley inter-regional network in the context of the Italian National Network for rare

- diseases. *Blood Transfus* 2014 Apr;12(Suppl 3):s617-s620 [FREE Full text] [doi: 10.2450/2014.0055-14s] [Medline: 24922303]
24. Palladini G, Kyle RA, Larson DR, Thorneau TM, Merlini G, Gertz MA. Multicentre versus single centre approach to rare diseases: the model of systemic light chain amyloidosis. *Amyloid* 2005 Jun;12(2):120-126. [doi: 10.1080/13506120500107055] [Medline: 16011989]
25. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. *J Am Med Inform Assoc* 2015 Jan;22(1):76-85 [FREE Full text] [doi: 10.1136/amiajnl-2014-002794] [Medline: 25038198]
26. APEC Rare Disease Network. Asia-Pacific Economic Cooperation. URL: <https://www.apec.org/rarediseases> [accessed 2021-01-08]
27. Nathan N, Taam R, Epaud R, Delacourt C, Deschildre A, Reix P, French RespiRare Group. A national internet-linked based database for pediatric interstitial lung diseases: the French network. *Orphanet J Rare Dis* 2012 Jun 15;7:40 [FREE Full text] [doi: 10.1186/1750-1172-7-40] [Medline: 22704798]
28. Griese M, Haug M, Brasch F, Freihorst A, Lohse P, von Kries R, et al. Incidence and classification of pediatric diffuse parenchymal lung diseases in Germany. *Orphanet J Rare Dis* 2009 Dec 12;4:26 [FREE Full text] [doi:10.1186/1750-1172-4-26] [Medline: 20003372]
29. Lavery A, Jaffé A, Cunningham S. Establishment of a web-based registry for rare (orphan) pediatric lung diseases in the United Kingdom: the BPOLD registry. *Pediatr Pulmonol* 2008 May;43(5):451-456. [doi: 10.1002/ppul.20783] [Medline: 18383113]
30. Bellgard M, Snelling T, McGree J. RD-RAP: beyond rare disease patient registries, devising a comprehensive data and analytic framework. *Orphanet J Rare Dis* 2019 Jul 12;14(1):176 [FREE Full text] [doi: 10.1186/s13023-019-1139-9] [Medline: 31300021]
31. Groft SC, Gopal-Srivastava R, Dellon ES, Gupta SK. How to Advance Research, Education, and Training in the Study of Rare Diseases. *Gastroenterology* 2019 Oct;157(4):917-921. [doi: 10.1053/j.gastro.2019.08.010]
32. Kaufmann P, Pariser A, Austin C. From scientific discovery to treatments for rare diseases - the view from the National Center for Advancing Translational Sciences - Office of Rare Diseases Research. *Orphanet J Rare Dis* 2018 Nov 06;13(1):196 [FREE Full text] [doi: 10.1186/s13023-018-0936-x] [Medline: 30400963]
33. Rubinstein Y, Groft S, Bartek R, Brown K, Christensen R, Collier E, et al. Creating a global rare disease patient registry linked to a rare diseases biorepository database: Rare

- Disease-HUB (RD-HUB). *Contemp Clin Trials* 2010 Sep;31(5):394-404 [FREE Full text] [doi: 10.1016/j.cct.2010.06.007] [Medline: 20609392]
34. Thompson R, Johnston L, Taruscio D, Monaco L, Bérout C, Gut I, et al. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J Gen Intern Med* 2014 Aug;29(Suppl 3):S780-S787 [FREE Full text] [doi: 10.1007/s11606-014-2908-8] [Medline: 25029978]
35. Kodra Y, Posada de la Paz M, Coi A, Santoro M, Bianchi F, Ahmed F, et al. Data Quality in Rare Diseases Registries. *Adv Exp Med Biol* 2017;1031:149-164. [doi: 10.1007/978-3-319-67144-4\_8] [Medline: 29214570]
36. Winther O, Svenstrup D, Henningsen PP, Kristiásson R, Jørgensen HL. FindZebra - the search engine for difficult medical cases. *Orphanet J Rare Dis* 2014;9(Suppl 1):O5. [doi: 10.1186/1750-1172-9-s1-o5]
37. Svenstrup D, Jørgensen HL, Winther O. Rare disease diagnosis: A review of web search, social media and large-scale data-mining approaches. *Rare Dis* 2015;3(1):e1083145 [FREE Full text] [doi: 10.1080/21675511.2015.1083145] [Medline: 26442199]
38. Pavitt K. What makes basic research economically useful? *Research Policy* 1991 Apr;20(2):109-119. [doi: 10.1016/0048-7333(91)90074-z]
39. Roll-Hansen N, London School of Economics and Political Science Centre for Philosophy of Natural and Social Sciences. Why the distinction between basic (theoretical) and applied (practical) research is important in the politics of science. In: Technical Report No. 04/09. London, United Kingdom: London School of Economics and Political Science; 2009.
40. Providing health intelligence to meet local needs: a practical guide to serving local and urban communities through public health observatories. WHO Centre for Health Development. 2014. URL: [https://apps.who.int/iris/bitstream/handle/10665/152645/9789241508162\\_eng.pdf?sequence=1&isAllowed=y](https://apps.who.int/iris/bitstream/handle/10665/152645/9789241508162_eng.pdf?sequence=1&isAllowed=y) [accessed 2021-01-08]
41. Doenças raras: o que são, causas, tratamento, diagnóstico e prevenção. Ministério da Saúde. URL: <https://tinyurl.com/y5vx8avf> [accessed 2021-01-12]
42. Pereira JDS, Machado WCA. Implantação de centro especializado em reabilitação: vantagens e desvantagens apontadas pelos gestores municipais de saúde. *Revista de Terapia Ocupacional da Universidade de São Paulo* 2015 Dec 26;26(3):373-381 [FREE Full text] [doi: 10.11606/issn.2238-6149.v26i3p373-381]
43. Medici A. Hospitais universitários: passado, presente e futuro. *Revista da Associação Médica Brasileira* 2001 Jun;47(2):149-156 [FREE Full text] [doi: 10.1590/s0104-42302001000200034]

44. Manual de Normas Técnicas e Rotinas Operacionais do Programa Nacional de Triagem Neonatal. Ministério da Saúde. 2002. URL: [https://bvsmms.saude.gov.br/bvs/publicacoes/triagem\\_neonatal.pdf](https://bvsmms.saude.gov.br/bvs/publicacoes/triagem_neonatal.pdf) [accessed 2021-01-12]
45. Conselho Nacional de Desenvolvimento Científico e Tecnológico. Chamada CNPq/MS/SCTIE/DECIT No 25/2019 - Inquérito sobre perfil de doenças raras no Brasil Internet. Ministério da Saúde. 2019 Aug 30. URL: <http://resultado.cnpq.br/6200770662786573> [accessed 2021-01-12]
46. Ministério da Saúde, lança a Rede Nacional de dados em Saúde e DATASUS realiza encontro técnico. Ministério da Saúde. URL: <https://datasus.saude.gov.br/ministerio-da-saude-lanca-a-rede-nacional-de-dados-em-saude-e-datasus-realiza-encontro-tecnico/> [accessed 2021-01-12]
47. ConecteSUS. Ministério da Saúde. 2019. URL: <https://conectesus-paciente.saude.gov.br/menu/home> [accessed 2021-01-12]
48. Bernardi F, Lima V, Pellison F, de Azevedo Marques PM, Rijo R, Galliez R, et al. Blockchain Based Network for Tuberculosis: A Data Sharing Initiative in Brazil. *Stud Health Technol Inform* 2019 Jul 04;262:264-267. [doi: 10.3233/SHTI190069] [Medline: 31349318]
49. Presidência da República. Lei No. 13.709, de 14 de agosto de 2018. Secretaria-Geral. 2018 Aug 14. URL: [http://www.planalto.gov.br/ccivil\\_03/\\_ato2015-2018/2018/lei/L13709.htm](http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm) [accessed 2021-01-12]
50. LGPD exigirá das instituições de saúde um gerenciamento cuidadoso das informações de pacientes. Portal Setor Saúde. 2019. URL: <https://setorsaude.com.br/lgpd-exige-das-instituicoes-de-saude-um-gerenciamento-cuidadoso-das-informacoes-dos-pacientes/> [accessed 2021-01-12]
51. Ebert C, Gallardo G, Hernantes J, Serrano N. *DevOps*. *IEEE Software* 2016 May;33(3):94-100. [doi: 10.1109/ms.2016.68]
52. Johnson HA. *Trello*. *JMLA* 2017 Apr 04;105(2):375. [doi: 10.5195/jmla.2017.49]
53. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. *Commun ACM* 1996 Nov;39(11):27-34 [FREE Full text] [doi: 10.1145/240455.240464]
54. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health* 2014 May 14;11(5):5170-5207 [FREE Full text] [doi: 10.3390/ijerph110505170] [Medline: 24830450]

55. World Health Organization. Data quality review: a toolkit for facility data quality assessment. Module 1: Framework and metrics. Geneva, Switzerland: World Health Organization; 2017.
56. Portaria No. 199, de 30 de janeiro de 2014. Ministério da Saúde. 2014 Jan 30. URL: [https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199\\_30\\_01\\_2014.html](https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html) [accessed 2021-01-12]
57. Portaria No. 981, de 21 de maio de 2014. Ministério da Saúde. 2014 May 21. URL: [http://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0981\\_21\\_05\\_2014.html](http://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0981_21_05_2014.html) [accessed 2021-01-12]
58. Bhardwaj S, Jain L, Jain S. Cloud computing: A study of infrastructure as a service (IAAS). *Int J Inf Technol Web Eng* 2010;2(1):60-63 [FREE Full text]
59. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
60. Harvard Humanitarian Initiative. KoBoToolbox: Data Collection Tools for Challenging Environments. KoBo Toolbox. 2018. URL: <https://www.kobotoolbox.org/> [accessed 2021-01-12]
61. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform* 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
62. Orphadata Ontologies. Orphadata. URL: <http://www.orphadata.org/cgi-bin/index.php#ontologies> [accessed 2021-01-12]
63. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002 Jan 01;30(1):52-55 [FREE Full text] [doi: 10.1093/nar/30.1.52] [Medline: 11752252]
64. Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier S, van Enckevort D, et al. Recommendations for Improving the Quality of Rare Disease Registries. *Int J Environ Res Public Health* 2018 Aug 03;15(8):1644 [FREE Full text] [doi: 10.3390/ijerph15081644] [Medline: 30081484]

65. World health statistics 2018: monitoring health for the SDGs, sustainable development goals. World Health Organization. 2018. URL: <https://apps.who.int/iris/handle/10665/272596> [accessed 2021-01-08]
66. Pessotti H, Mazzer G, Barbosa JF, Alves D. Portal ORAH: Ferramentas para exploração de informações de internações hospitalares. 2012 Presented at: XIII Congresso Brasileiro de Informática em Saúde; Nov 19-23, 2012; Curitiba, Brazil.
67. Souza J, Alves D. New Web Application Allows Healthcare Decision Makers to Monitor the Performance of Hospitals through Inpatient Quality Indicators. 2015 Presented at: 6th European Conference of the International Federation for Medical and Biological Engineering; Sep 7-11, 2015; Dubrovnik, Croatia. [doi: 10.1007/978-3-319-11128-5\_182]
68. Rigoli F, Mascarenhas S, Alves D, Canelas T, Duarte G. Tracking pregnant women displacements in Sao Paulo, Brazil: a complex systems approach to regionalization through the emergence of patterns. *BMC Med* 2019 Oct 01;17(1):184. [doi: 10.1186/s12916-019-1416-4] [Medline: 31570106]
69. Carvalho I, Lima V, Yazlle Rocha JS, Alves D. A tool to support the clinical decision based on risk of death in hospital admissions. *Procedia Computer Science* 2019;164:573-580. [doi: 10.1016/j.procs.2019.12.222]
70. Gattini C. Implementing National Health Observatories: operational approach and strategic recommendations. Pan American Health Organization. 2009. URL: <https://tinyurl.com/y5xrapan> [accessed 2021-01-08]
71. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One* 2017;12(1):e0170365 [FREE Full text] [doi: 10.1371/journal.pone.0170365] [Medline: 28099516]
72. Aymé S, Bellet B, Rath A. Rare diseases in ICD11: making rare diseases visible in health information systems through appropriate coding. *Orphanet J Rare Dis* 2015 Mar 26;10:35 [FREE Full text] [doi: 10.1186/s13023-015-0251-8] [Medline: 25887186]
73. Salvadores M, Alexander P, Musen M, Noy N. BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. *Semant Web* 2013;4(3):277-284 [FREE Full text] [Medline: 25214827]
74. Pinnock H, Epiphaniou E, Sheikh A, Griffiths C, Eldridge S, Craig P, et al. Developing standards for reporting implementation studies of complex interventions (StaRI): a systematic review and e-Delphi. *Implement Sci* 2015 Mar 30;10:42 [FREE Full text] [doi: 10.1186/s13012-015-0235-z] [Medline: 25888928]

75. Theobald S, Brandes N, Gyapong M, El-Saharty S, Proctor E, Diaz T, et al. Implementation research: new imperatives and opportunities in global health. *Lancet* 2018 Nov 17;392(10160):2214-2228. [doi: 10.1016/S0140-6736(18)32205-0] [Medline: 30314860]
76. Strategy 2018-2023: building the science of solutions. World Health Organization. 2017. URL: <https://apps.who.int/iris/handle/10665/255777?locale=zh> [accessed 2021-01-08]
77. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health* 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
78. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM Planning and Evaluation Framework: Adapting to New Science and Practice With a 20-Year Review. *Front Public Health* 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]

## 7 A PROPOSAL FOR A SET OF ATTRIBUTES RELEVANT FOR WEB PORTAL DATA QUALITY: THE BRAZILIAN RARE DISEASE NETWORK CASE

### 7.1 Introduction

A web portal is a web-based repository of data, information, facts, results of analyzes, and knowledge [1]. Its main objective is to provide an active working environment for the user, that, in the context of this work, can be a researcher, policy maker, manager, or healthcare professional, among other stakeholders [2]. According to the World Health Organization (WHO), in the health domain, these portals also can be understood as health observatories as long as they gather information relevant to a given health phenomenon of a population, with the potential to generate population indicators capable of supporting public health policies [3].

Web portals in health stand out due to the isolated and non-interoperable character that data in this area tend to present [2] and the potential of such portals to mitigate these problems [3]. Still, repositories of epidemiological data are useful and contribute to the planning of public health programs [4]. Also, a web portal is the most effective and efficient form of dissemination and communication, allowing the generation of knowledge and intelligence in various areas of health [2].

Given the relevance of health web portals, great care is required with the information contained therein. Stakeholders must accept such information as valid, objective and scientifically sound. Also, it must be useful and relevant to health policies, actions, plans, programs, and projects as they must be valid, reliable, coherent, representative, sensitive, understanding, and ethical [5].

However, when managers, researchers, and professionals in the health area try to identify the population's health needs and measure, evaluate and monitor the impact of health actions and policies, they often encounter relevant issues in health data [3]. Such a problem is the deficiency in accessibility, lack of quality [6] and the non- interoperable character of data [2] in this area.

Deepening the discussion for the specific case of Rare Disease (RD) in Brazil, their data also show problems. In this case, after the institution of the Brazilian Policy for Comprehensive Care for People with Rare Diseases by the Unified Health System ('SUS'; Portuguese: 'Sistema Único de Saúde') in 2014 [7], the lack of structured data on RD in Brazil became evident. To mitigate this problem, the Brazilian Network of Rare Diseases ('RARAS'; Portuguese: 'Rede Nacional de Doenças Raras') project was established in 2020 aiming to conduct a national



survey regarding the clinical overview, epidemiology, diagnostic and therapeutic resources, and costs for RD of genetic and non-genetic origin in Brazil and to create a national surveillance network of RD [8].

The project has its funding from the Ministry of Health of Brazil through the National Council for Scientific and Technological Development of Brazil and encompasses many health centers that carry out RD diagnosis and treatment across all five Brazilian regions [8].

In addition, the RARAS survey had two phases, the first being a retrospective data analysis of patients who attended RDs centers between the years 2018 and 2019 and the second being a prospective analysis of the patients assisted in 2022, between April and September [8]. The standardized electronic Case Report Forms (eCRFs) were designed in Research Electronic Data Capture (REDCap) to collect data from the health centers in order to structure the collection and analysis processes and to provide data quality monitoring [9].

In this context, the literature highlights the relevance of validating a health web portal. Despite widespread growth in the use and complexity of web portals, insufficient attention is paid to their quality. The purpose of this paper is to describe the validation process of the RARAS Portal and identify a set of data quality attributes required for psychometric evaluations that will support the portal implementation.

## **7.2 Related Work**

RDs represent a complex public health problem worldwide. The multidisciplinary required in this area makes the task of establishing local and isolated information solutions impracticable. In this context, according to the WHO, to overcome these barriers, the research and development in the RDs must involve collaboration networks, and this can be done through health web portals and observatories, which generate indicators capable of helping the public health policies [10].

Thus, since the 1990s, important initiatives have been developed to meet these demands. In 1997, the French National Institute for Health and Medical Research (INSERM) established the first scope of the Orphanet platform, a unique web service that gathers and produces knowledge about RDs, to improve procedures related to diagnosis, care, and treatments of people with RDs [11].

The Orphanet (<https://www.orpha.net/>) offers validated and quality information with easy access. In addition, maintaining and supporting standardized terminology for RDs (ORPHA code) is an essential tool to highlight and categorize the intrinsic and specific characteristics of RDs, improving their visibility and interoperability in health information

systems. This initiative became a continental project in Europe from 2000 onwards. Currently, 41 European countries contribute to this project [11]. This platform also offers information services to align development and research processes in this area, such as Orphanet Rare Disease Ontology (ORDO), a structured dictionary for RDs, and Orphadata, a web resource with data related to orphan drugs [12].

After that, other initiatives started to emerge. The PhenomeCentral (<https://phenomecentral.org/>) is a web portal created to enable secure sharing of the recording of phenotypic descriptions of patients with RDs. It is still possible to identify clinical characteristics associated with genes in the database of this web service. PhenomeCentral incorporates data from over 1,000 patients with rare genetic diseases and has already been used to identify RD-causing mutations in many patients. EpigenCentral (<https://epigen.ccm.sickkids.ca/>), is a free web resource for biomedical researchers, molecular diagnostic laboratories, and clinicians to perform the classification and interactive analysis of DNA data related to RDs. It allows users to search for patterns associated with known diseases in their data and classify the pathogenicity of genetic variants to aid molecular diagnosis or analysis patterns [13].

All these projects use, at some level, a method of evaluating their functionality. In this paper, we will describe scientifically validated methods and guidelines for this type of evaluation and the model for their application in our web portal, the RARAS Portal.

## **7.3 Methods**

### *7.3.1 Study design and participants*

This protocol describes a cross-sectional study of mixed nature divided into three steps that will occur asynchronously and independently. The first phase will involve participants from centers eligible for electronic data gathering of the RARAS study and who are indirectly engaged in the construction of the portal, including coordinators, fellows, volunteers, supervisors, and focal points (local officials). For the second phase, specialists from universities in all regions of Brazil and servers/technicians from the Ministry of Health will be invited to participate in three distinct moments in the portal evaluation process. The third phase will involve institutions and patients' associations, as well as members of civil society and the pharmaceutical industry that will see the final version of the portal for the first time after being validated by members and specialists.

The proposed protocol is composed of a set of instruments and steps. The adoption of the instruments follows the criteria: (1) instruments must be based on recognized international organizations' standards, good practice guidelines, and quality data principles; (2) instruments used in several other studies published in renowned publications; and (3) the instruments must assess desirable psychometric qualities such as reliability, validity, and sensitivity. The adoption of the steps for the present protocol follows the following criteria: (1) the steps should allow the use of the protocol in different health research areas; (2) the protocol should include all procedures of the selected instruments; and (3) the protocol should indicate how to use the results of the application of each instrument. To achieve this task, we perform a brief literature review to identify the existing key works to evaluate the main process, metrics, and guidelines for assessing a health web portal quality.

### *7.3.2 Evaluation Process, Metrics/Measures and Guidelines*

Based on a thorough literature review, the selected instruments are the key renowned instruments that met the criteria indicated in the literature. Therefore, we chose to evaluate usability in the first stage; user experience and consensus in the second stage, and e-Service quality in the third. These aspects are critical components of software quality, but they are often overlooked during the development stage and can hamper the results in software products with poor usability, user experience, and quality of service.

#### First Stage: Usability

According to NormanGroup, a world-leading research-based User Experience (UX) organization, usability is one of the most important factors in the Human-Computer Interaction (HCI) field. Nielsen defines usability in five dimensions: learnability; efficiency; memorability; few errors; and user satisfaction [14]. This classical definition or a usability framework is the most widely adopted and cited since it provides a detailed articulation of usability aspects that can be objectively and empirically verified through different evaluation methods.

There is a need to determine what constitutes usability in terms of its components or dimensions and how it can be evaluated. Without this understanding, it is difficult to consider usability during software development or perform appropriate software usability evaluation. Because of that, we will also proceed with a qualitative evaluation through empathy mapping

[15]. We intend this technique to aid in decision making, exploring the user attitudes and behaviors, and reveal any holes in existing user needs.

#### Second stage: Delphi consensus

A Delphi consensus is a structured process used to evaluate expert opinions on health and medical topics. It uses a series of questionnaires that are iterated until a consensus is reached. Although it does not rely on statistical power but on group dynamic to reach a consensus among experts, some specialists' features are desirable as education and field of expertise, accumulated experience, and willingness to engage in the research, in addition to a recommended sample size of between 5 and 20 experts in the Delphi panel. This method was chosen as the most suitable research method to address the research objective, as the results of the previous literature review study will inform the initial statements for the first round of the research [16].

Throughout all the three rounds, respondents will be allowed to elaborate on their decisions and recommend additional or alternative domains and measurement instruments. Participants who did not complete one of the three rounds will be excluded from further participation.

Experts who agree to participate will receive a personal email invitation containing an anonymous web link to the first Delphi round. In this round, participants will be asked to indicate whether every proposed domain was important for the outcome measurement of RD web portals' quality attributes, with three response options ('yes', 'no' and 'this is not my area of expertise'). If the reply is affirmative, the respondent will be asked to indicate whether the measurement instruments proposed are suitable for the concerned domain ('yes', 'no' and 'no opinion'). Thus, the first round will aim to identify potentially suitable measurement instruments for the RD web portals' quality attributes. Hence, experts will not be asked to consider a preference for a particular measurement instrument until the second round but rather were asked to indicate whether a proposed instrument would be suitable for RD web portals.

Based on the consensus obtained in the first round, we intend to verify, in the second round, if the same instruments are applicable and would cover beyond their proposed scope of minimal attributes for international RD portals web [17]. Thus, the identification of the core domains will be carried out in the second round. To narrow down the selection of key domains, we will rank the domains that reached consensus in the first-round ascending by the percentage of 'Yes' responses. Regarding the domains for which there was no consensus on the level of the measurement instrument, participants will be asked to indicate their preferences, placing the

proposed instruments in order of preference (first place is most preferred, last is least preferred). Finally, the third round will enclose questions to help clarify the last issues regarding overlap between selected attributes presented at the RARAS Portal.

### Third Stage: eService Quality

A web portal is also responsible for many consultations of medical and health information every day. Most of its users only analyze the information on the first ten sites retrieved in a search [18]. In this sense, understanding the critical factors to improve a website's traffic through Search Engine Optimization (SEO) is essential to assess the UX and the adhesion and engagement of new users, which can be boosted by the reach and quality of a website [19]. Even if the concept of UX is very broad and complex, a good UX results from the obtained evidence, mostly qualitative. One such evidence is electronic service quality (e-SQ) monitoring.

The e-SQ is defined as the extent to which services based on web technology facilitate effective and efficient online communications, purchases and delivery of products or services. Its concept includes five dimensions: reliability, assurance, tangibles, empathy, and responsiveness. In terms of attributes, this corresponds to technical adequacy, content, security, communication, prestige, ease of use, ease of learning, memorability, layout, graphics, system availability, speed, accessibility, navigation, reliability, accuracy, privacy, contact information, online help, responsiveness, sustainability and currency [18]. As indicated, many of these e-service quality dimensions and attributes can also be applied to usability and UX, measured by stages 1 and 2, respectively.

Thus, one part of e-SQ is planning and optimizing web projects. However, it also means monitoring website performances. For performance assessment, we choose off-the-page tools that check single performance indicators and especially the visibility rank of hosted websites represents a very effective approach. These SEO tools (described in section 3.3), will provide digital marketing metrics like the number of total organic and their goal completions via organic traffic, the bounce rate of top landing pages, top exit pages, target keywords and their organic ranking, mobile usability, and crawled errors under search console [20]. To find out the engagement interest and adoption of our portal from the perspective of users, a questionnaire survey will be conducted. The questionnaire survey will aim to find out the opinions, attitudes, and satisfaction of users with the selected e-services provided by the portal and measure the lead conversion rate.

### *7.3.3 Instruments and tools*

REDCap is a metadata-driven application built at Vanderbilt University in 2004 to enhance clinical research. The software is free and widely used by the scientific community to collect and manage research data. It enables classical and translational clinical research, basic science research, and general surveys, providing researchers with a tool for the design and development of data collection instruments in a flexible way [21]. For the survey portion of this study, several decisions have been made to ensure that data collection is carried out ethically. When participants enroll in the survey portion of the study, they will have an opportunity to take as much time as they need to read an electronic study information letter and ask questions of the research team before beginning the survey. In this sense, the REDCap modules of alerts, data quality, and electronic consent (e-consent) should be used.

The voluntary nature of the study will be communicated in this letter. Once participants have begun to fill in the survey, they can decide to stop at any point without penalty. Participant responses to the survey will be collected via an online survey platform. All data in the online survey are stored on a secure server at the study site, enhancing participant data security. Regarding the specialist panels portion of the study, informed consent will be obtained prior to the first-round beginning.

For the first stage to assess the impact of the portal on healthcare professionals and organizations, we propose using Computer System Usability Questionnaire (CSUQ). CSUQ was validated with 825 employees who worked at nine IBM development sites [22]. Also, for assessing the portal usability from the technical perspective we will use the adapted model proposed by Yoshiura [2]. This model describes opportunities for developing and implementing new observatories or for the adequacy of existing Health Observatories through components based on information technology multi-layer architecture [23]. The starting point for the elaboration of the script of questions for the second stage will be a review of the literature on the minimum set of data and information [24] and the FAIR (Findability, Accessibility, Interoperability, and Reuse) principles of digital assets [25]. For the second phase of this stage, a set of preliminary portals will be presented, and finally, for the third phase, our portal will be submitted for expert evaluation.

Finally, for the third stage, monitoring tools such as Google Analytics and Hotjar will be used. These tools may help web designers to better understand how users behave on their websites, such as how they interact with a product or feature. The tools are also useful for detecting bugs and discovering usability issues. For instance, opportunities for improvements can arise by analyzing quantitative data from Google Analytics, combined with qualitative data from Hotjar [26]. An e-SQ questionnaire based on Ssemugabi [27] will also be available for

voluntary participation and will be present in the outreach campaign in non-participating institutions.

#### *7.3.4 Data Analysis Pipeline*

Computing the RARAS Portal's efficiency is essential to minimize the risk that the user does not return to the service and experience good navigation on the website. Regarding the first stage of Section 3.2, the questionnaires contain closed questions with possible answers as strongly disagree; disagree; neutral; agree; totally agree. Although qualitative, each response will be assigned to a quantity according to the participant's level of satisfaction (0 to 4). Therefore, statistics over the contentment of the first stage can be computed for each one of the categories: purpose, interface, usability, content, and communicational aspects. Then, several metrics can be calculated from the rate of each answered question.

One important metric is computed by the two-way analysis of variance (ANOVA). Unlike the classical one-way metric, the two-factor statistics can help the technical team evaluate average grades over each group of questions and categories, i.e., two independent variables. The two-way ANOVA allows us to understand the variance of average grades within and between each question category, i.e., two independent variables. The analysis of the statistical metric will allow the technical team to understand and further investigate the quality of the RARAS Portal based on the initial value of the F-test [28].

The same methodology will be applied for each round of the Delphi study with the assignment of quantitative points such as totally agree (4) when there is a consensus. Another important test to evaluate users' experience surveys is the student's t-Test [29], which can evaluate the responses of different types of users (medical staff and general users of the website) or over different periods. Furthermore, it is essential to calculate the standard deviation (std) for each one of answered questions in the usability analysis. This standard deviation will allow us to understand the divergence between user experience, which can be further evaluated according to the results of the Delphi study, the exploitation of users' Hotjar heatmap, and Google Analytics.

For stage 3, some of the evaluated metrics obtained by Google Analytics and Hotjar will be the number of clicks on a single page [30] as well as the heat map of a page, the percentage of usage on a page, the amount of traffic and their source (social media, google, among others), time of response of the webpage, cursor distance of different tasks [31]. All metrics will

compose a dashboard to help the information technology team in the continuous improvements of the RARAS Portal even after the first two stages, usability and Delphi study.

### *7.3.5 General Evaluation Framework: RE-AIM*

The RE-AIM model has been applied at various stages of evaluative research. Its application has helped many researchers and managers in the planning and evaluating programs, both at the individual (target population) and organizational (program provider) levels, thus seeking to reduce the gaps between research and practice and maximize the impact of public health interventions. The model is composed of five dimensions: Reach (number, proportion, and representativeness of people who could or want to participate in an intervention); Effectiveness (the impact that a given intervention has on outcomes, including negative effects); Adoption (number, proportion, and representativeness of organizations and environments that are willing to adopt it); Implementation (fidelity and consistency to the intervention protocol); and Maintenance (long-term effects of a program at the individual level and institutionalization of the intervention at the organizational level) [32].

The results of the usability, consensus building, and service quality stages will be organized into RE-AIM dimensions. Reach indicators will include lead conversion rate, participant characteristics, and focus group participation rates. Effectiveness will include results of consensus on the presence of attributes and mechanisms capable of making the RARAS Portal properly locatable and accessible, capable of interoperating with external agents and being reused when requested. Qualitative feedback obtained through empathy mapping of the resources and delivery potentials that the portal has will be used as indicators of Adoption. For a better understanding of this dimension, keywords, traffic objectives, and heat maps will also be considered.

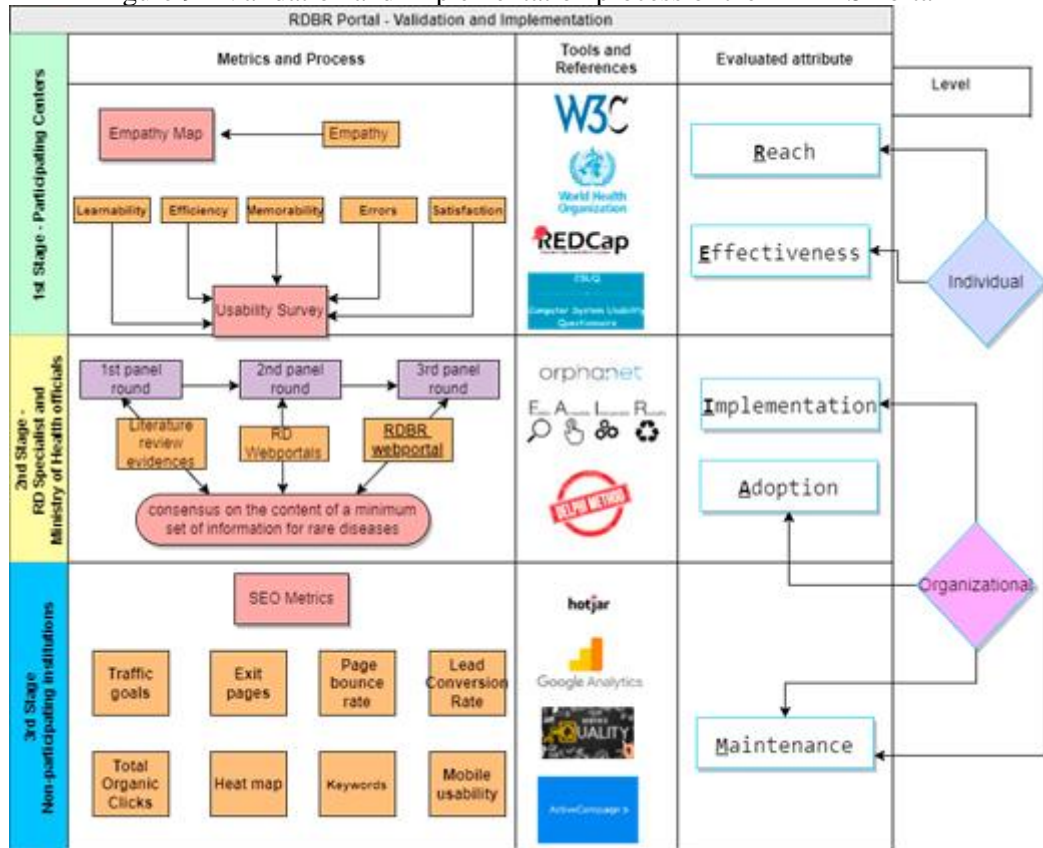
Implementation indicators will include reporting of technical difficulties, the number of days of activity recorded, and participants' perceptions of the portal. Maintenance was divided into individual and organizational. The individual maintenance indicators included quantitative and qualitative results on the usability of each monitor. Experts' perceptions of the intervention will indicate potential organizational maintenance.



### 7.4 Related Work

We hope that the systematic development methods and the validation process of the RARAS Portal allow the success of the technological tool developed and reproduce a scientific character to the development process, differentiating it from other existing health sites. In summary, the improvement carried out during the validation of RARAS can contribute to the dissemination of knowledge in the area and include, based on scientific knowledge and clinical experts, offering clear, attractive and accessible information for the population. Therefore, we hope that the RARAS Portal will become a widespread, useful and reliable source of health information, especially for RD patients and professionals, in a quality interactive environment. Figure 9 presents the interaction between the protocol stages and the RE-AIM framework dimensions and levels.

Figure 9 - Validation and Implementation process of the RARAS Portal



The results may be used to guide a broad understanding of the aspects involved in the process and gaps in the implementation of health portals. This understanding can improve their use and adoption and then support better planning and dissemination of RD situations at the national level according to each region's particularity.

Furthermore, the information gathered in this study can provide subsidies for better decision-making in the area of RD and also for planning and creating health policies.

Our proposal has some limitations: Service quality for websites is not a fully defined construct and there is still uncertainty in defining and interpreting its meaning, mainly because there are various types of sites. As such, there is no established conceptual model for developing and evaluating the service quality of websites in general. Besides that, our evaluation analyses just a determined period, it's desirable to understand how the portal scalability and maintenance work for a long time.

The RE-AIM framework can be used to understand how technology can increase user engagement on the RARAS Portal, identifying what stage of the implementation process the various digital technologies are currently at. It can also highlight areas where more research may be needed or processes in place to move from research to implementation.

## 7.5 Final considerations

The protocol proposed in this work may result in the first Brazilian portal of centralized and reliable information on RD. To the best of our knowledge, no previous studies have been made to comprehensively model and evaluate the e-service quality, usability, and user experience of RD web portals in a consolidated manner.

Therefore, a research gap exists in determining components of integrated e-service quality, usability, and user experience evaluation model in general, and for RD web portals, in particular. We also expect to evaluate and provide a web portal as easy to use, where the main focus is usability, and enjoyable to use, with the main focus on UX. In addition, users of the RARAS Portal need to get validated content and the required reliable online services without doing exhaustive searches or visiting multiple sources, with the main focus on e-service quality.

## 7.6 References

- [1] Rodrigues, R. J., & Gattini, C. H. (2017). "National Health Information Systems and Health Observatories." In *Global Health Informatics* (pp. 14-49). Academic Press.
- [2] Yoshiura, V. T. (2020). "Desenvolvimento de um modelo de observatório de saúde baseado na web semântica: o caso da rede de atenção psicossocial" (Doctoral dissertation, Universidade de São Paulo).
- [3] World Health Organization. (2016). "Guide for the establishment of health observatories." Available at: <https://apps.who.int/iris/handle/10665/246123,%20last%20accessed%202022/06/30>
- [4] Alves, D., Yamada, D. B., Bernardi, F. A., Carvalho, I., Colombo Filho, M. E., Neiva, M. B., ... & Félix, T. M. (2021). "Mapping, infrastructure, and data analysis for the Brazilian Network of Rare Diseases: protocol for the RARASnet Observational Cohort Study." *JMIR Research Protocols*, **10**(1), e24826.

- [5] Gattini, C. H. (2009). "Implementing National Health Observatories: operational approach and strategic recommendations." *Implementing National Health Observatories: operational approach and strategic recommendations*.
- [6] Pereira, B. D. S., & Tomasi, E. (2016). "Instrumento de apoio à gestão regional de saúde para monitoramento de indicadores de saúde." *Epidemiologia e Serviços de Saúde*, **25**, 411-418.
- [7] Brasil. Portaria nº 199, de 30 de Janeiro de 2014, Ministério da Saúde. Available at: [https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199\\_30\\_01\\_2014.html](https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html), last accessed 2022/02/16.
- [8] Félix, T. M., de Oliveira, B. M., Artifon, M., Carvalho, I., Bernardi, F. A., Schwartz, I. V., ... & Alves, D. (2022). "Epidemiology of rare diseases in Brazil: protocol of the Brazilian Rare Diseases Network (RARAS-BRDN)." *Orphanet Journal of Rare Diseases*, **17**(1), 1-13.
- [9] Yamada, D. B., Bernardi, F. A., Neiva, M. B., Lima, V. C., Vinci, A. L. T., de Oliveira, B. M., ... & Alves, D. (2022). "National Network for Rare Diseases in Brazil: The Computational Infrastructure and Preliminary Results." In *International Conference on Computational Science* (pp. 43-49). Springer, Cham.
- [10] Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E., & Cornel, M. C. (2008). "Orphanet: a European database for rare diseases." *Nederlands tijdschrift voor geneeskunde*, **152**(9), 518-519.
- [11] de la Paz, M. P. "Improved Diagnosis and Care for Rare Diseases through Implementation of Precision Public Health Framework." *Rare Diseases Epidemiology: Update and Overview*, **55**.
- [12] Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., ... & Rath, A. (2020). "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database." *European Journal of Human Genetics*, **28**(2), 165-173.
- [13] Turinsky, A. L., Choufani, S., Lu, K., Liu, D., Mashouri, P., Min, D., ... & Brudno, M. (2020). EpigenCentral: Portal for DNA methylation data analysis and classification in rare diseases. *Human Mutation*, **41**(10), 1722-1733.
- [14] Nielsen, J. (2012). "Usability 101: Introduction to usability." NNGroup. Available at: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- [15] Nielsen, J. (2012). "Empathy Mapping: The First Step in Design Thinking." NNGroup. Available at: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- [16] Haughey, D. (2010). Delphi technique a step-by-step guide. Project Samart. co. uk.
- [17] Domensino, A. F., Winkens, I., van Haastregt, J., van Bennekom, C. A., & van Heugten, C. M. (2020). "Defining the content of a minimal dataset for acquired brain injury using a Delphi procedure." *Health and quality of life outcomes*, **18**(1), 1-10.
- [18] Rayess, H. M., Gupta, A., Nissan, M., Carron, M. A., & Zuliani, G. F. (2017). Search engine optimization: an analysis of rhinoplasty web sites. *Facial Plastic Surgery*, **33**(06), 665-669.
- [19] Pilarcikova, K., Rusnak, P., Rabcan, J., & Kostolny, J. (2019, November). User experience in the development of the education system. In *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)* (pp. 626-632). IEEE.
- [20] Lewoniewski, W., Härting, R. C., Węcel, K., Reichstein, C., & Abramowicz, W. (2018, October). "Application of SEO metrics to determine the quality of Wikipedia articles and their sources." In *International Conference on Information and Software Technologies* (pp. 139-152). Springer, Cham.
- [21] Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). "Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support." *Journal of biomedical informatics*, **42**(2), 377-381.
- [22] Lewis, J. R. (1995). "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use." *International Journal of Human-Computer Interaction*, **7**(1), 57-78.
- [23] Yoshiura, V. T., Yamada, D. B., Pellison, F. C., De Lima, I. B., Damian, I. P. M., Rijo, R. P. C. L., ... & Alves, D. (2018). "Towards a health observatory conceptual model based on the semantic web." *Procedia computer science*, **138**, 131-136.
- [24] Bernardi, F. A., Yamada, D. B., de Oliveira, B. M., Lima, V. C., Félix, T. M., & Alves, D. (2022). "The minimum dataset for rare diseases in Brazil: a systematic review protocol." *Procedia Computer Science*, **196**, 439-444.
- [25] Lamprecht, A. L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., ... & Capella-Gutierrez, S. (2020). "Towards FAIR principles for research software." *Data Science*, **3**(1), 37-59.
- [26] Kierkegaard, E. (2021). "Optimizing the sign-up flow for a fintech company using Google Analytics, Hotjar and A/Btesting."
- [27] Ssemugabi, S. (2019). "Development and Validation of an Integrated Model for Evaluating E-Service Quality, Usability and User Experience (E-SQUUX) of Web-Based Applications in the Context of a University Web Portal." Unpublished PhDthesis.
- [28] Armstrong, R. A., Eperjesi, F., & Gilmartin, B. (2002). "The application of analysis of variance (ANOVA) to different experimental designs in optometry." *Ophthalmic and Physiological Optics*, **22**(3), 248-256.
- [29] Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). "Application of student's t-test, analysis of variance, and

covariance." *Annals of cardiac anaesthesia*, **22**(4), 407.

[30] Ferreira, J. M., Acuña, S. T., Dieste, O., Vegas, S., Santos, A., Rodríguez, F., & Juristo, N. (2020). "Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction." *Information and Software Technology*, **117**, 106195.

[31] Harrati, N., Bouchrika, I., Tari, A., & Ladjailia, A. (2016). "Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis." *Computers in Human Behavior*, **61**, 463-471

## 8 FROM RAW DATA TO FAIR DATA: THE FAIRIFICATION WORKFLOW FOR BRAZILIAN TUBERCULOSIS RESEARCH

### 8.1 Introduction

Although tuberculosis (TB) is entirely curable and preventable, it affects individuals who lack sustainable living conditions and easy access to information. In Brazil, patients with TB infections who do not have private health plans can be treated by public hospital healthcare systems through the Unified Health System (SUS). The SUS has established a series of programs specifically aimed at controlling, dining, and supporting patients infected with TB. One is the National Program for TB Control, which aims to eliminate TB as a public health problem in Brazil. However, challenges must still be overcome to achieve this objective [1].

Most notably, there is a need to create, adapt, and introduce new diagnostic and treatment technologies and contribute to research on innovative technologies. Strengthening control and information systems (IS) for the disease in vulnerable populations, who are the most affected, is also necessary. Activities such as public health program control and organization, health facility management, ensuring the availability of medical supplies, and the lack of instruments may hinder the quality of care from supporting the definition of management parameters [2].

Among the main factors that negatively influence the decision-making process, it is possible to highlight the low quality, availability, and integration of population health data. Several computerized systems are created to follow up information about patients with TB in practical and quick ways. These systems prioritize collecting data such as personal information from patients, medications, treatments, diagnostic tests, and control of routines. The IS addressed in this work were: SISTB (a local IS), TBWEB – Tuberculosis Patient Control System of São Paulo State; SINAN – Information System for Notifiable Diseases and GAL – Laboratory Environment Manager, both are national systems, but for data accessibility reasons, we used samples from the state of Rio de Janeiro. Despite the existence of such data, some reasons make it difficult for managers and healthcare professionals to access them, such as the lack of digitization of processes, heterogeneity, and duplication of data in health IS, and the existence of a large amount of isolated data accessible only in a particular context [3].

To address these difficulties, such as the lack of expansion of scientific discoveries and the discovery of new knowledge, low communication between obtained data, and non-easily reproducible data, the Global Open FAIR (Findable, Accessible, Interoperable, and Reusable) initiative was created. It aims to disseminate a large-scale recommendation that supports better

management of research data in open science, access to research data, and scientific information by machines and humans. The FAIR methodology is a solution for standardizing data and sharing information about the disease [4]. So, this study highlights the difficulty in conducting research based on TB data available in Brazil.

## 8.2 Methods

Due to the information heterogeneity in the systems above, the FAIR principles can be considered weak and ambiguous enough to lead to different interpretations. Thus, it is necessary to clarify its meaning and define criteria to evaluate the data FAIRification process. Since the maturity of FAIR data depends on the capabilities of the ecosystem components, the FAIR assessment should include not only FAIR data as an outcome but also a specific assessment of criteria relevant to each essential element of every process involved in the flow of information from systems, taking into account the combination of data and services they [5]. To calculate the metrics centered on the data present in the TB IS, that is, on the variables they collected, the dimensions described by Bernardi et al. (2022) [6] and the instrument validated by the European project “Fostering FAIR Data Practices in Europe” (<https://fairaware.dans.knaw.nl/>).

## 8.3 Results and Discussion

The quality and relevance of the information produced to know the population's health conditions may be compromised when there are variables with inadequate completion. Incomplete data make it impossible to assess other quality dimensions and use techniques that allow the crossing of information. Juxtaposed, overcoming such challenges can improve the filling quality and expand the scope of use of this information in epidemiological studies and decision-making. The SATIFYD questionnaire was designed to understand the compliance of the datasets evaluated against the FAIR principles. In this sense, the higher the score, the more synchronized the dataset is with the FAIR Guidelines. Table 10 describes the TB IS analyzed and their compliance with the FAIR principles.

**Table 10** - Brazilian Tuberculosis Information Systems (BTIS) compliance with the FAIR principles.

BTIS	Informational Level	Total Records	Timeframe (2000's)	FAIR Element %	Total FAIR %
TBWEB	State (São Paulo)	208.624	06 - 19	F:55 A:50 I:25 R: 22	38%
SINAN	National (State Sample)	104.541	01 - 14	F:27 A:50 I:25 R: 22	31%
GAL	National ((State Sample)	11.651	10 - 17	F:66 A:50 I:25 R: 22	41%
SISTB	Municipal (Ribeirão Preto)	4.065	00 - 22	F:66 A:50 I:25 R: 22	41%

The validation and adequacy of filling out collection instruments and databases of TB IS requires monitoring and evaluating completeness, which helps identify data weaknesses and strengths and recommend strategies to improve information quality. Completeness is still a little-explored quality dimension in TB IS in Brazil, and data often needs to be standardized or transformed. Compliance check is commonly implemented from pre-established guides and reference tables by organizations to ensure conformance to policies and standards.

These standards ensure that tasks are performed correctly and provide order within the organization. Complying with the same standards will prevent many errors, such as data duplication, capturing incorrect spellings, and using inaccurate formats. Despite this, the absence of standards observed in TB IS makes it impossible to compare their databases. The impossibility of reaching databases also reveals a barrier to data reliability, represented by precision. Precision is the degree to which the data show the truth about the described event. The sine qua non condition for the existence of accurate information is preceded by complete and correctly represented data.

Coding errors and poor documentation are the primary reasons for data non-validation in information systems. While post-submission curation can improve data quality retrospectively, implementing practical data management solutions early in project design is crucial. FAIR principles, which promote better data availability and reuse, address interoperability and harmonization issues with health information systems. These requirements aim to improve clinical decision-making and are summarized by Kodra et al. [7] in their perspective on 'quality informatics' using health data collected from IS.

This study suggests that studies on improving data quality in information systems can help identify problems such as inadequate data collection instruments, lack of researcher training, and technical document review. Strategies that will enhance data quality, including integrated databases, can enable the recovery of incomplete or inconsistent data in Brazilian IS for TB. The study notes the use of metrics to support data evaluation according to the FAIR principles, where a compliance rate of 37.75% was achieved.

These findings indicate no significant gap between awareness of the need to implement the FAIR guidelines in health IS and the incorporation of data quality promotion activities based on the FAIR guidelines. Therefore, all the main actors involved, including those who generate data and administrators of IS, should be encouraged to know their strengths and weaknesses. Continuously fostering strategies to promote data quality is a potent stimulus for strengthening national health IS and can potentially benefit from recommendations on how to overcome the inherent limitations of these IS.

The FAIRfication process for tuberculosis data in Brazil faces data quality, privacy issues, access difficulties, technological limitations, and different data sources. Future work to address these limitations includes improving data quality, developing ethical guidelines for data availability, enhancing technological infrastructure, standardizing data formats, and investing in research. These initiatives can ensure interoperability, availability, and reuse of tuberculosis data and promote disease control.

#### 8.4 Conclusion

Data quality management in the Brazilian TB IS is not yet systematically carried out. The evaluation of only some parts of the information production cycle relies solely on specific and independent metrics. This necessitates further research in this field to develop systematic methods that consider the specific characteristics of each computerized scenario and identify their potential contributions to improving the quality of TB information.

#### 8.5 References

- [1] Do Carmo IA, Maia JC, De Novaes JV, Almeida Lde, Pereira NA, da Costa GV, et al. Os Desafios para o controle da tuberculose no Brasil. *Brazilian Journal of Health Review*. 2022;5(6):23969–78. doi: 10.34119/bjhrv5n6-168
- [2] Makeleni N, Cilliers L. Critical success factors to improve data quality of electronic medical records in Public Healthcare Institutions. *SA Journal of Information Management*. 2021;23(1). doi: 10.4102/sajim.v23i1.1230.
- [3] Thapa C, Camtepe S. Precision Health Data: Requirements, challenges and existing techniques for data security and privacy. *Computers in Biology and Medicine*. 2021;129:104130. doi: 10.1016/j.compbiomed.2020.104130
- [4] Courtot M, Gupta D, Liyanage I, Xu F, Burdett T. BioSamples database: Fairer samples metadata to accelerate research data management. *Nucleic Acids Research*. 2021;50(D1). doi: 10.1093/nar/gkab1046
- [5] Jacobsen A, Kaliyaperumal R, da Silva Santos LO, Mons B, Schultes E, Roos M, et al. A generic workflow for the data fairification process. *Data Intelligence*. 2020;2(1-2):56–65. doi: 10.1162/dint\_a\_00028
- [6] Bernardi FA, Alves D, Crepaldi NY, Yamada DB, Lima VC, Rijo RPCL. Data Quality in health research: an integrative literature review. *medRxiv*, 2022-05. doi: 10.2196/preprints.41446.
- [7] Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier S, van Enckevort D, et al. Recommendations for improving the quality of rare disease registries. *International Journal of Environmental Research and Public Health*. 2018;15(8):1644. doi: 10.3390/ijerph15081644.



## 9 A COMPUTATIONAL INFRASTRUCTURE FOR ANALYZING TUBERCULOSIS RESEARCH DATA IN BRAZIL

### 9.1 Introduction

Tuberculosis (TB) is the second highest cause of death caused by a single infectious agent, surpassing HIV/AIDS. Despite being a curable and preventable disease, in 2021, 6.4 million new cases were recorded, resulting in 1.4 million deaths among HIV- negative individuals and 187,000 deaths among HIV-positive individuals [1]. The emergence of drug-resistant strains can be attributed to low-quality medication, poor hygiene, use of inappropriate medication, and delayed treatment approaches [2].

Drug-resistant TB, which is expensive and requires a long treatment period, needs new, affordable, and effective diagnostic tools that guarantee quality and proven efficacy to be rapidly implemented [3]. In addition, health services should incorporate new information systems to aid decision-making. The World Health Organization launched the Stop TB Strategy in 2006, followed by a new strategy in 2015 with more ambitious goals and a greater focus on research and innovation. Although both approaches have significantly reduced TB cases in high-burden countries, multidrug- resistant tuberculosis remains a global problem [4].

Data reliability is critical for improving health service quality and creating effective public policies. However, generating health information is faced with barriers such as problems with patient data documentation, data interpretation difficulties, and organizational issues [5]. In TB research, a large amount of complex data is produced from dispersed sources with low integration and varying accuracy levels. It impedes knowledge extraction and data analysis, making it challenging to provide decision support in operational and administrative processes and scientific research [6].

Currently, data coordination between TB stakeholders parties can be messy, prone to delays, subject to manipulation, and obscure. Thus, this work aims to describe a computational pipeline and the infrastructure needed for analyzing TB research data, assisting in establishing a high-quality data source in Brazil.

### 9.2 Methods

#### *9.2.1 Brazilian Tuberculosis Research Network Ecosystem*

The most used diagnostic methods for TB consider bacteriological, radiological, histopathological, and immunological approaches. The bacteriological tests consist of

bacilloscopy and culture. Clinical materials such as sputum, bronchial and bronchoalveolar lavage, and other samples that can be taken from the respiratory tract are used for TB research [2].

The clinical laboratory has a fundamental role in the health system, given that most medical decisions are made using the information provided by laboratory processes. Quality assurance in a clinical analysis laboratory is built on all process stages, from the material collection (pre-analytical) to the result (post-analytical) delivery. Clinical samples sent to the laboratory for TB diagnosis must comply with a series of general conditions on which the quality and efficiency of the test results depend. It is essential to control the data quality from local centers belonging to the Brazilian TB research network, which covers 65 institutions and researchers [3].

#### *9.2.2 Data Gathering, Infrastructure, Curation, and Analysis Pipeline*

The TB Network conducts studies that require the collection and management of data in scientific, clinical, and managerial/epidemiological domains. The Research Electronic Data Capture (REDCap) platform is used for this purpose, which is a web-based application that allows the creation of case report forms, surveys, and research databases. REDCap is an open-source application that provides several tools for exporting data in multiple formats, including the CSV standard that facilitates compatibility with other third-party tools.

The collected data are categorized and segmented inside the network through the projects, and participants have restricted access based on the network policies. Data are automatically anonymized to ensure ethical, legal, and confidential issues, and their confidence rate level can be chosen. Before analysis, the data undergoes pre-processing to identify and treat missing and abnormal values and duplicate and redundant data. Validations are also carried out to ensure the types of variables and transformations, such as normalization and discretization.

The exploratory analysis is the next step in the analysis pipeline, where the data's distribution and amount are better understood. Descriptive statistics such as the mean, median, and standard deviation are calculated. Graphics such as bar and scatter charts and visual elements such as presentations and images are used to understand the results better. Regression and clustering calculations can also be used to find patterns and relationships in the data that are not easily identifiable.

### 9.2.3 Auxiliary Tools Gathering, Infrastructure, Curation, and Analysis Pipeline

The collaboration scripts for developing machine learning and artificial intelligence algorithms are developed in Google Colab, a Jupyter Notebook-based platform. It provides an accessible and readily available environment for developing Python programs. Python is popular due to its high-level nature and simplicity, which enables portability across different platforms, and it has a vast developer community and numerous libraries. Specific Python libraries used in programming tasks include Numpy for mathematical functions such as linear algebra, Pandas for data analytics functions, and Scikit-learn for machine learning algorithms. Streamlit, a Python-based data visualization platform that allows data transformation into shareable pages without prior knowledge of other programming languages, was adopted for data visualization. Streamlit can be easily integrated with Python applications, requiring minimal adaptation to receive Python outputs, and is an efficient way to display data without using large tables.

### 9.2.4 TBWeb Application for TB Analysis

A research network developed a web application to validate statistical analyses related to project data. The portal can connect to any database and offers real-time statistical analysis of clinical data through data visualization techniques. Researchers can monitor all updates in the data in real time.

## 9.3 Expected Outcomes

The expected primary outcomes of this work are tools and a curated knowledge basis for frameworks that can promote clinical data quality within the Brazilian tuberculosis stakeholders through a national research network. Over the medium time, we expect to also encourage new models of data sharing (e.g., safe data havens, data lakes, data hubs) and innovative privacy-preserving and processes analytical methods. Also, it is expected, in the long term, to obtain positive health outcomes, such as developments in public health indicators, a better understanding of health services processes, improved research outcomes, and new approaches to ethical and legal issues.

Although the final goal is to implement this pipeline to the national network in Brazil, in phase 1, it will be first deployed and validated in 7 research centers, leading to different data types. Data from several government sources will be gathered to build a test database for the

computational tool and validate data sharing among peers. Our success indicators will be based on the following parameters: adoption of the tool by the TB Network entities; the volume of data available and used; and perception of usefulness for academic, clinical, and managerial audiences. After defining these parameters, a validation process for an official implementation as a national tool for TB will be carried out with the TB experts independently committee.

Based on a well-established process and robust evidence, we hope we can compare our results with the source documents of each center. It will allow checking the validity of a sample of the data entered on the form and define intervention, such as visiting the local centers to improve the data culture and promote continuous education. We expect also to be enabled to compare our solution with consolidated approaches like the Observational Health Data Sciences and Informatics (OHDSI) tools and the FAIR Principles.

## 9.4 Conclusions

It is expected to successfully build a high-quality data source to provide a basis for developing new decision-support tools. We hope to advance scientific research and establish new diagnosis algorithms and optimized operational models toward better patient care and managerial decisions. In the long term, it is expected to achieve positive health outcomes such as improved public health indicators, a better understanding of health service processes, and new approaches to ethical and legal issues.

## 9.5 References

- [1] Global tuberculosis report 2022 [Internet]. World Health Organization. World Health Organization; [cited 2023Mar28]. Available from: <https://www.who.int/publications-detail-redirect/9789240061729>
- [2] Kanabalan RD, Lee LJ, Lee TY, Chong PP, Hassan L, Ismail R, Chin VK. Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. *Microbiological research*. 2021 May 1;246:126674. Available at: <https://doi.org/10.1016/j.micres.2020.126674>.
- [3] Kritski A, Andrade KB, Galliez RM, Maciel EL, Cordeiro-Santos M, Miranda SS, Villa TS, Ruffino Netto A, Arakaki-Sanchéz D, Croda J. Tuberculosis: renewed challenge in Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*. 2018 Jan;51:02-6. Available at: <https://doi.org/10.1590/0037-8682-0349-2017>.
- [4] Uplekar M, Raviglione M. Who's end TB strategy: From stopping to ending the global TB epidemic. *Indian Journal of Tuberculosis*. 2015;62(4):196–9.
- [5] Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. *BMC health services research*. 2017 Dec;17:1-0. Available at: <https://doi.org/10.1186/s12913-017-2697-y>.
- [6] Bernardi FA, Alves D, Crepaldi NY, Yamada DB, Lima VC, Rijo RP. Data Quality in health research: an integrative literature review. *medRxiv*. 2022:2022-05. doi:10.1101/2022.05.31.22275804.

## 10 PROPOSAL FOR A HEALTH INFORMATION MANAGEMENT MODEL BASED ON LEAN THINKING

### 10.1 Introduction

Planning, reviewing processes, monitoring performance, and constant improvements within organizations have become fundamental in recent years. The quality systems were created to promote competitiveness, efficiency, effectiveness with high-performance indexes of the institutions, obtaining successful results [1]. The implementation of quality management in health institutions is an arduous task, which requires efforts from different stakeholders that involve management, such as people and information management, material resources, among others, becoming the angular stone for can seek continuous improvement of health management processes and procedures, through the implementation of routines in the workplace and changes in the behavior of employees [2].

Managing quality in public health institutions is a challenge, not only for the manager or due to the resources available to meet expenses, but also due to the need to restructure the production chain, eliminating care models that may disorganize the organizational culture reflecting on quality in the service provided [3]. Lean thinking, also known as lean thinking and Toyota Production System (STP), means doing more with less, less time, less space, less human effort, less use of machinery, less material, offering customers what they need. they need [4]. Waste or change is any process that requires resources and does not add value, that is, they need corrective measures, are products and services that the customer does not want and do not meet their real need, high stocks, unnecessary processes, movement of employees, transportation of unnecessary products and employees who are waiting for some activity that was not carried out within the deadline to follow production [5].

For the implementation of Lean to be successful within the institution, it is essential that managers play an active role in several areas, such as adopting planned methodology in relation to implementation; provide necessary resources; designate those responsible for the process; distribute responsibilities and involve the team; to emphasize the importance of teamwork together (well-developed communication channels; ascending and descending); manage expectations (i.e. fear of losing your job); ensure that employees understand the need for change (i.e. new roles according to the change implemented); creating an experimentation environment, with a culture of understanding risk and a safety net for trials and failures; make the team understand about competitive Lean reasons, such as benefits for the organization and for those involved in the process, presenting the future's look after the change; analyze and

share information regarding cost-benefit; delegate and emphasize the responsibility of each one [6].

## **10.2 Background**

### *10.2.1 Related Work*

In the information and communication era, different areas of health are adopting new and diverse tools and techniques for the development of better solutions aiming to guarantee the quality, delivery time, sustainability, cost reduction, and even greater organizational performance in a scenario notoriously characterized by the marked production of information. In contrast, the availability and knowledge generated in these environments are mostly evidenced by the scarcity of appropriate infrastructure and effective integrative approaches [7]. Thus, the organization and structuring of information in a synthesized way and with the purpose of identifying and formalizing their relationships, conceive knowledge, a fundamental principle that forms the basis of numerous studies in several areas [8].

Data processing is essential in decision-making processes and in assessing the quality of health services. The management of health systems requires mechanisms capable of dealing with administrative aspects that represent the conditions of organization and functioning of the various levels that make up health services. The management of a health service involves taking care of the organizational and functional aspects associated with it. In addition, the health administration process must be composed of information systems that process data related to the individual's health and life condition, in addition to the environmental conditions and other factors that interfere in the health-disease process [9].

Although assessing quality in the health field is a relevant challenge, there is unanimity among managers that it is essential to select appropriate assessment systems and methods to assist the administration of services and provide decision-making with the least degree of uncertainty possible. Many institutions use quality indicators as an instrument to measure quality, with the aim of identifying how and where improvement can be made. Therefore, health indicators are useful to assess and monitor the activities carried out by health services, in addition to contributing to the identification of the degree of risk of the occurrence of a certain event or health problem, as well as checking values and acquiring information that allows intervening in the reality you want to know, in order to achieve goals and objectives [10].

Lean, also known as lean philosophy, is a management model that has been used in the area of Health [11], and has its origin in the automobile industry. Different applications in common problems in the routine of many health services such as long queues, rising costs, and various types of recurring waste such as inventory, administration, and logistics were solved through small actions based on the Lean methodology. The ability to increase the agility and documentation of processes, reduce errors and indirect costs, and optimize the use of resources are presented as the main benefits when applied to the operational context. In addition, studies suggest Lean as a strategy that can favor a profound transformation at the organizational and managerial levels [12].

The use of Lean in health aims to provide an increasingly accurate service to patients, with cost reduction and that does not cause harm from the time they enter until they leave the health organization; solving problems of greater occurrence within the sector, such as long lines, high costs and recurring waste [13]. This thinking in health is guided by six principles: Lean is to create value; Lean is an attitude of continuous improvement; Lean is a unit of purpose respect for the people who develop the work; it's visual; is standardization with flexibility [14].

The principles of Lean employed in health are positive and add improvements for several reasons, as health organizations are divided into departments (silos) and, often, the only person who sees the patient's flow is himself. In these systems, the path of the patient is composed of long periods in different health institutions, so that the value of the aggregated information is summarized in small intervals of time. When applying this thinking, specifically in the context of the continuous flow of value, it is essential that the deconstruction of processes is done. Thus, in a scenario where the patient provides systems with dispersed information, it is possible to ensure changes that occur across functional boundaries [15]. Lean Healthcare promotes patient-focused assistance, in which therapeutic and interventionist results offer improvements in operational management, with the satisfaction of the technical team and patients, reducing waste and costs.

### *10.2.2 Goals*

The main goal is to propose a health information management model based on Lean thinking that acts as an auditable instrument of analysis, representation, and improvement of the quality of health information in the municipality of Ituverava through medical procedures in the different areas that make up the Unified Health System (SUS). As specific objectives of this work, we have:

- review the different models, thoughts, and philosophies applied to health information systems, in order to support the proposed methodology based on the best practices exposed in the literature;
- evaluate the different information, flows and practices present in the health information systems that make up the location of the present study;
- know the levels of flow bottlenecks within a hierarchical and regionalized system in order to identify the waste (seedlings) of information present in the municipality;
- improving the reliability of the information, allowing managers to have greater visibility through the use of visual tools and providing greater security in the decision-making process;
- provide means for using audit filters to assess the quality of care provided to patients to identify strengths and weaknesses, allowing corrective actions.

In the next section, some key concepts for the execution of this article will be presented. The fourth section presents the research proposal with the methodology that will lead the proposed study and the discussion in the fifth section. Finally, the sixth section summarizes the conclusions and next steps for the development of a more comprehensive model on a management model of health information systems based on the Lean philosophy in the municipality of Ituverava / SP.

### **10.3 Research Protocol**

#### *10.3.1 Study design*

According to Vergara [16], research on the management theme can be classified as to the ends and the means. As for the purposes, it refers to applied, exploratory, and methodological research. As for the means, this work included bibliographic and documentary research in the stage of characterization of health information systems, their processes, in addition to a mapping of those responsible and users of these systems, present in the municipality of Ituverava. In this sense, the activities will be planned and developed in three phases:

- Phase 1 - bibliographic and documentary study: Integrated by the conceptual framework and the identification of precedent elements for the development of the management model. In line with one of the objectives of this work, this phase includes three activities:
  - a) survey of the portfolio of information systems that support the activities of the



municipality and the context in which these applications are used; b) bibliographic review in order to identify the state of the art of using Lean tools in relation to health information management; c) mapping and characterization of the processes involved in the information routine at different levels of health care.

- Phase 2 - Model structuring and information management: Includes the description of the execution steps and the complete specification of the management model. The first activity aims to specify, based on the theoretical framework, the architecture of the model, which included the quality model with its dimensions and peculiarities, the instruments for sensitive assessment for each dimension, in addition to the elaboration of a map in order to identify those responsible, users and other stakeholders involved with the systems. The second activity of this phase will consist of the selection, analysis, and mapping of the data flow and application of value to the processes explained in Phase 1 for the structure of the management model.
- Phase 3 - Evaluation of the applicability and effectiveness of the model: in this step, activities will be developed to verify the effectiveness of the management model. The first activity will review the processes in order to assist decision making based on the literature review and the intervention tools to be applied in each process. In the end, we will have a report on the processes for health information systems, identifying possible waste (seedlings) of information, allowing the improvement of the application of resources, and, therefore, making a more effective intervention choice.

### *10.3.2 Data source*

Public health operation and management processes are supported by an information technology infrastructure, through a set of nationwide information systems, including epidemiological, primary care, outpatient and hospital events, and among other actions carried out by the Ministry of Health. The data sources included in this study will come from health information systems that integrate the dimensions of information related to the municipality of Ituverava at government, state, and regional levels.

### *10.3.3 Literature Review*

An integrative literature review will be carried out to map the different techniques and processes applied to this context, this stage of the study is a Scoping Review study and will be prepared according to the methodology of the Joanna Briggs Institute (JBI).

The search will be performed electronically in the databases: Latin American and Caribbean Health Sciences Literature (LILACS), Web of Science, National Library of Medicine (PubMed), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Scielo, ScienceDirect, SCOPUS and World Health Organization Library Database (WHOLIS). For the combination of descriptors, the Boolean terms will be considered: AND, OR, and NOT composing the search formulas.

After conducting the search, it is estimated to include research conducted in English and Portuguese, with a quantitative and qualitative approach, primary studies, systematic reviews, meta-analyses and/or meta-syntheses, books and guides, published or not published until the present period. and answer the question of the established search. Websites and advertisements in the media will be excluded. For the search, descriptors and their synonyms will be used according to the Health Sciences Descriptors (DeCS) and Medical Subject Headings (MeSH).

### *10.3.2 Lean Tools*

Lean thinking consists of principles and techniques, the first being associated with the institution's philosophy, which are the bases that guide lean strategic actions such as establishing value for the customer, defining flow, maintaining continuous flow, pulled production, integration of the supply chain, focus on quality, visual management, use of technology, technical staff and processes, development of human resources and continuous improvement. The second is the means by which the principles are achieved and maintained, namely: value stream mapping (VSM), just in time, kanban, automation (jidoka), five Ss (5S), standardization, workload leveling (heijunka), group technology and cell layout, employee according to takt time, zero-defect quality control, total production maintenance, visual control, multi-professional and teamwork, empowerment ( autonomy) and kaizen [17].

The tools used in lean thinking are mechanisms for applying and structuring the results for eliminating waste and adding value throughout the process. Tools such as Value Stream Mapping, 5S, Heijunka, Single Minute Exchange of Die (SMED) or quick tool change, Poka-Yoke, Kanban, Kaizen, Visual Management, Standardized Work, Gemba, Andon and Total Productive Maintenance (TPM). These tools should help to identify the processes that do not

add value to the organization so that the results can be optimized and acquire a competitive advantage [18].

#### **10.4 Discussion**

Interinstitutional data production and sharing have grown considerably in recent years. According to Miloslavskaya and Tolstoy [19], humanity has generated more data in the past two years than in its entire past history. However, there is a great difficulty for institutions to analyze the data produced, due to its volume, speed, veracity, variety, and values found in the different sectors that integrate health services [20]. Due to the complexity, scope, and particularities of the SUS, the management of data, knowledge, and health services must be carefully performed, so that quality care can be offered at all levels of care. In addition, the diversity of forms, structures, and patterns present in information systems represents a challenge for the safe and reliable exchange of information [21].

In this sense, the definition of organized processes and the adoption of continuous thinking in the improvement of the activities performed are extremely important to overcome the heterogeneity and intangibility present in the spheres that constitute health information systems (HIS). The heterogeneity of the data is intrinsically connected to the type of information generated by health services. The highly heterogeneous, and sometimes ambiguous, nature of the medical language and its constant evolution, the high amount of data generated constantly by the automation of processes and the emergence of new technologies and the need to process, analyze and make decisions based on this information constitute the foundation for the inevitable computerization of health to promote the production and management of knowledge [22].

The essence of Lean thinking is the elimination of superfluous activities, that is, waste, which interposes itself in the most diverse processes, assistance, support, and administrative. By eliminating effort that does not add value, there will be time and resources accessible for things that are needed. Eliminating waste means doing what is relevant, providing more space to improve the quality of work and patient safety; so that the processes to actions become fast, efficient and effective, while at the same time reducing costs and improving the workplace [23].

When inserting Lean thinking in a health organization, managers first need to assess the organizational structure, as the vast majority of health care organizations are structured at hierarchical levels with vertical decision-making movement, that is, from the top to the bottom. base. Lean has a better performance when in a horizontal process, going from a need for service until its delivery, without ceasing the flow; in order to make it possible, leaders must organize

their employees into operational teams, orienting themselves on the individual products or services offered to the patient. Such lack means that the team needs to understand that they work for the patient and not for the sectors of the institution, which demands a reorganization of the team to satisfy the requirements of the process [24].

Among the main aspects that negatively influence the decision-making process, we can highlight the low quality, availability, and integration of population health data. Despite the existence of such data, there are reasons that prevent access by managers and health professionals, such as the non-computerization of processes, the heterogeneity, and duplicity of data in health information systems, and the existence of a large amount of isolated data and accessible only in a given context. Thus, health data is commonly found dispersed through independent systems and fragmented in closed databases. Such factors can cause problems of quality of the information, making it difficult to coordinate and evaluate them, considering that, despite the intense volume, the information remains decentralized without it is possible to assist the decision-making process [25].

Lean Information Technology (IT), also called Lean IT can be defined as the participation of the team through the use of Lean principles, with systems and tools, incorporating, aligning and synchronizing IT management with the area of business; being able to offer quality information with an effective information system, promoting continuous improvement and innovation of processes[26].

For the implementation of Lean IT to have effective results, it is necessary for the manager to be able to: (i) have a vision of what is important for the management of the institution and choose Lean tools that are appropriate for IT complications; (ii) integrate the implementation of the tools together with five management components: strategy, processes, structure, performance and culture metrics. Thus, when implementing Lean strategies in IT, it is necessary to evaluate all its processes within the institution to eliminate waste, structure functions within the applied methodology, and measure improvement at all levels of the organization, promoting the Lean culture [27].

According to Paesa [28], the principles are the first accepted truths, and after them, the laws constitute the theory. The method establishes that the steps must be adopted, with autonomy represented by the choice of technology to be used. Taking the method, choosing the technology, and specifying its form of application, we have the technique itself. Among the elements that are fundamental to the application of Lean, the principles evaluate this philosophy as continuous improvement applied to the institution's processes. The techniques are listed as operational, for evaluating their practice in daily use. In this way, it is possible to move with

the institution's team between a high level of abstraction (Lean principles) and a level of practical application (Lean techniques). A recurrent cause of failure in Lean implementations is the lack of understanding of Lean as a philosophy [29], [30].

Techniques without principles lead to the blind application, oblivious to decisive factors. Only through the principles is it possible to apply the techniques. The principles are more comprehensive in their recommendations, with a degree of abstraction that hinders their operationalization, due to this factor it is necessary to resort to techniques [28].

### **10.5 Conclusions and future work**

The implementation of Lean thinking in health opens up a range of opportunities for continuous improvement, because through the integration with the team on concepts and tools it is possible to build a health system with low costs, with less waste, reducing waiting time and queues, adding value in each process, without overloading the team through changes in habits, that is able to understand the real need of the user and able to deliver what he really needs in a timely manner. Increasing resources in a health organization, whether inputs, technology, finances, and human when the whole process is unstructured, will generate more waste, affecting the institution, the team, and especially the customer. Adopting Lean thinking in health may seem a challenge initially for managers and team members, but as the first results begin to appear, profound and concrete changes are visible for positive transformation for improvements in the quality of the service provided, until the culture can be learned completely in order to have the perfect care.

### **10.6 References**

- [1] BONATO, V. L. Gestão de qualidade em saúde: melhorando assistência ao cliente. **Revista O Mundo da Saúde**, São Paulo, v. 35, n.5, p. 319, 2011.
- [2] SAKODA, T. J. **Gestão de qualidade na saúde**. 2011. 57f. Trabalho de Conclusão de Curso (Graduação) – Gestão de Projetos, Universidade Presbiteriana Mackenzie, São Paulo, 2011.
- [3] SAVASSI, L. C. M. Qualidade em serviços públicos: os desafios da atenção primária. **Revista Brasileira de Medicina de Família e Comunidade**, Rio de Janeiro, v. 7, n.23, p. 71, 2012.
- [4] DENNIS, P. **Produção Lean Simplificada: um guia para entender o sistema de produção mais poderoso do mundo**. 2.ed. Porto Alegre: Bookman, 2008. p. 31-32.
- [5] PRATES, C. C.; BANDEIRA, D. L. Aumento de Eficiência por Meio do Mapeamento do Fluxo de Produção e Aplicação do Índice de Rendimento Operacional Global no Processo Produtivo de uma Empresa de Componentes Eletrônicos. **Gestão & Produção**, São Carlos, v. 18, n. 4, p. 705-706, 2011.
- [6] CUNHA, A. M. C. A.; CAMPOS, C. E.; RIFARACHI, H. H. C. Aplicabilidade da metodologia Lean em uma lavanderia hospitalar. **Revista O Mundo da Saúde**, São Paulo, v.35, n. 5, p. 312-313, 2011.

- [7] MAKADY, A.; HAM, R. T.; DE BOER, A.; HILEGE, H.; KLUNGEL, O.; GOETTSCHE, W. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. **Value in Health**, v. 20, n. 4, p. 520-532, 2017.
- [8] RONQUILLO, C.; CURRIE, L. M.; RODNEY, P. The evolution of data-information-knowledge-wisdom in nursing informatics. **Advances in Nursing Science**, v. 39, n. 1, p. 1-18, 2016.
- [9] CARVALHO, A. O.; EDUARDO, M. B. P. **Sistemas de informação em saúde para municípios**. In: *Sistemas de informação em saúde para municípios*. 1998.
- [10] DENTLER, K. **Computing healthcare quality indicators automatically: secondary use of patient data and semantic interoperability**. 2014. 220 f. Doctoral dissertation - VU University Amsterdam, Amsterdam, 2014.
- [11] DURUR, F.; AKBULUT, Y. Lean Methodology for Pathology Laboratories: A Case Study from a Public Hospital. **Turkish Journal of Pathology**, v. 1, n. 1, p. 1-9, 2019.
- [12] SHAH, R.; WARD, P.T. Lean manufacturing: context, practice bundles, and performance. **Journal of Operations Management**, v. 21, n. 2, p. 129-149, 2003.
- [13] PERALTA, C. B. L.; COUTO, B. S.; BORBA H. B. Propostas de Melhorias em um Processo de Posto de Saúde com Auxílio do Lean Healthcare. **Associação Brasileira de Engenharia de Produção**, Rio de Janeiro, p. 5, 2015.
- [14] PINTO, C. F. *Em busca do cuidado perfeito: aplicando Lean na saúde*. 1. ed. São Paulo: Lean Institute Brasil, 2014. p.10-135.
- [15] INACIO, B. C.; ARAGÃO, J. F.; BERGIANTE, N. C. R. Implementação da Metodologia Lean Healthcare no Brasil: um estudo bibliométrico. **Associação Brasileira de Engenharia de Produção**, p. 9, 2016.
- [16] VERGARA, S. C. **Projetos e relatórios de pesquisa**. São Paulo: Atlas, 2006.
- [17] REGIS, T. K. O.; GOHR, C. F.; SANTOS, L. C. Implementação do Lean Healthcare: experiências e lições aprendidas em hospitais brasileiros. **Revista de Administração de Empresas**, São Paulo, v. 58, n. 1, p. 31, 2018.
- [18] MOREIRA, S. **Aplicação das ferramentas lean: caso de estudo**. 2011. 113f. Dissertação (Mestrado) - Instituto Superior de Engenharia de Lisboa - ISEL, Lisboa, 2011
- [19] MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. **Procedia Computer Science**, v. 88, p. 300-305, 2016.
- [20] ANTONIOU, G.; BARYANNIS, G.; BATSAKIS, S.; GOVERNATORI, G.; ROBALTO, L.; SIRAGUSA, G.; TACHMAZIDIS, I. Legal reasoning and big data: opportunities and challenges. **Legal Reasoning and Big Data: Opportunities and Challenges**, 2018.
- [21] LIMA, C. R. A.; SCHRAMM, J. M. A.; COELI, C. M.; SILVA, M. E. M. Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. **Cadernos de Saúde Pública**, v. 25, n.10, p. 2095-2109, 2009.
- [22] KONOPKA, B. M. Biomedical ontologies- a review. **Biocybernetics and Biomedical Engineering**, v. 35, n. 2, p. 75-86, 2015.
- [23] PINTO, C. F.; BATTAGLIA, F. **Aplicando Lean na Saúde**. Lean Institute Brasil. Disponível em: <[https://www.lean.org.br/comunidade/artigos/pdf/artigo\\_262.pdf](https://www.lean.org.br/comunidade/artigos/pdf/artigo_262.pdf)>. Acesso em: 09 maio 2019.
- [24] JOINT COMMISSION RESOURCES. **O Pensamento Lean na Saúde: menos desperdício e filas e mais qualidade e segurança do paciente**. Porto Alegre: Bookman, 2013. p. 8-35.
- [25] BRASIL, Ministério da Saúde. *A experiência brasileira em sistemas de informação em saúde*. Série B. **Textos Básicos de Saúde**, v. 2, 2009.

- [26] CANTANHEDE, M. A. D. **Lean Thinking em Desenvolvimento de Software: estudo e aplicação de ferramenta para avaliação do lean em software.** 143 f. 2014. Dissertação (Mestrado em Tecnologia). Faculdade de Tecnologia da Universidade Estadual de Campinas. Campinas, 2014.
- [27] LOURENÇO, N. M. A. **Transformação Numa Área de Serviços TI e Implementação de Metodologia Lean.** 71 f. 2011. Projeto (Mestrado em Gestão). Instituto Universitário de Lisboa, Lisboa, 2011.
- [28] PASA, G. S. **Uma Abordagem para Avaliar a Consistência Teórica de Sistemas Produtivos.** 188 f. 2004. Tese (Doutorado). Universidade Federal do Rio Grande do Sul, 2004.
- [29] BHASIN, S.; BURCHER, P. Lean viewed as a philosophy. **Journal of Manufacturing Technology Management**, v. 17, 2006.
- [30] DAHLGAARD, J, J; DAHLGAARD-PARK, S. M. Lean Production, Six Sigma Quality, TQM nad Company Culture. *The TQM Magazine*, v. 18, n. 3, p. 263-281, 2006.

## 11 AN INFORMATION SYSTEM FOR MONITORING TUBERCULOSIS CASES: IMPLEMENTATION RESEARCH PROTOCOL USING RE-AIM FOR A HEALTH REGION IN BRAZIL

### 11.1 Introduction

Tuberculosis (TB) is a curable disease and, until the coron avirus (COVID-19) pandemic, it was the leading cause of death from a single infectious agent [1]. The treatment is considered long, lasting at least six months, and to ensure that the medications are taken correctly, the Directly Observed Treatment (DOT) is recommended, which consists of observing patients taking the medications at least during the intensive phase of treatment (the first two months) [2].

The global plan to fight TB, approved in 2015 at the World Health Assembly, may not meet the targets set for 2035 [3]. Years of progress in contributing services and reducing the burden of disease have been rolled back due to the COVID-19 pandemic. As a result of the pandemic, there has been a global drop in the number of people diagnosed with TB. While there were large increases in diagnoses between the years 2017 and 2019, there was an 18% drop between 2019 and 2020, from 7.1 million to 5.8 million diagnoses [1]. Global TB targets are mostly far from being met [1]. To achieve the goals, research and innovation were identified as one of the three essential pillars for ending the TB epidemic through discovery and equitable access to innovative tools and approaches at national and global levels [4].

In this scenario, digital technologies and computing methods for obtaining scientific results play a central role, such as helping patients and caregivers to improve the experience in health facilities, improve treatment processes and automate the management of health information [5,6]. Thus, research that considers operational aspects of the treatment and control of its dissemination are still needed. Considering this scenario, the need to develop software that provides real-time information about the patient's path along health information systems for fast and adequate clinical decision-making is fundamental. A system that includes information on TB cases in a unified database may provide a regional epidemiological scenario, assisting in the action plan to fight the disease.

In order to assist in the TB treatment, an information system called SISTB [7,8] was developed. The system was implemented in a municipality [8,9], but it is intended to expand its reach to a region composed of 26 municipalities. However, the implementation of a new technology can change the way of working and result in the creation of new roles and responsibilities, as well as requiring changes in existing ones [10]. It is necessary to understand the environments where digital technologies will be implemented to identify their potential role



in the various aspects of TB care. In this way, it is possible to use implementation research as an approach where it is possible to investigate the challenges and provide evidence to guide the correct use and expansion of digital technology in TB care [11].

Implementation research is an interface between tools, strategies, and interventions and their use in health systems [12]. It considers real-world conditions, providing the basis for context-specific decision-making, which is crucial to turning theory into practical reality [12, 13]. In implementation research, different sources and perspectives are explored, to describe the processes used in the implementation of initiatives, as well as the contextual causes that affect these processes, seeking to understand what is going right or wrong in the implementation [13]. Implementation research links intervention effectiveness with research on effective processes for incorporating interventions into existing settings, and the RE-AIM framework provides a model to inform this research [14]. The RE-AIM framework was developed to facilitate the translation of research into practice, considering both internal and external validity [14, 15]. It has five dimensions: Reach, Effectiveness, Adoption, Implementation, and Maintenance. They operate at the individual, staff, and setting levels and interact to determine the impact of an intervention on the population [15].

Therefore, this work reports a protocol to evaluate the implementation of a technological intervention, that is, a health information system to assist in the monitoring of TB treatment.

## **11.2 Objectives**

The first objective is to characterize the way TB treatment is handled in a region composed of 26 municipalities, with the following sub-objectives: to characterize the TB treatment support services in each municipality in the region; determine the referral flow of assistance from the municipalities; to identify the behavior of the flow of information on treatment between services and municipalities.

The second objective is to evaluate the implementation of the information system in TB services in the region using the RE-AIM framework as support, with the following sub-objectives: building a pre-implantation protocol considering the diversity of municipalities; evaluate the implementation of the system using RE-AIM; assess the intention to continue using the system.

The third objective is to identify the barriers and facilitators in the implementation and use of the system.

## **11.3 Methods**

### *11.3.1 Intervention*

The intervention of this study is to implement an information system to assist in monitoring TB treatment in health services in a region comprising 26 municipalities. The purpose of the SISTB is to replace the paper forms used in the treatment, especially the daily follow-up form of the DOT so that all information is accessed immediately after registration by different professionals and services [7-9].

Additionally, it is expected to centralize treatment data to provide better support for case investigation. In the region where the study will be carried out, different data on TB cases can be found in distinct systems, such as the Tuberculosis Case Monitoring and Notification System (TBWEB), the Notifiable Diseases Information System (SINAN), the Tuberculosis Special Treatments Information System (SITE-TB), the Laboratory Environment Manager (GAL), the National Primary Health Care Information System (eSUS-AB), and, finally, the electronic medical records of each municipality/service. Thus, SISTB will include an interoperability layer, making it possible to exchange data with other information systems used by the services to reduce duplicate registrations, rework for health professionals, and obtain information from as many sources as possible in a single system [7, 16 -18]. Access to the aforementioned systems will be requested to carry out the interoperation between them and the SISTB. In cases where interoperability is not possible, the data from that system will be requested to be used as a secondary source to obtain as much data as possible of treatment cases in all possible instances.

SISTB will also include an evidence-based decision support system and national protocols to monitor and analyze data collected or produced in the various TB patient care settings. Thus, it will be possible to incorporate data processing, analysis, and machine learning techniques to assess the clinical situation and probable patient risks in real time [7, 19, 20].

### *11.3.2 Study context*

The study context will be a health region composed of 26 municipalities, based in Ribeirão Preto, State of São Paulo, Brazil, with an estimated population of 1,400,000 inhabitants [21]. In this region, care in specialized services (secondary level of care) for TB is carried out in the four largest municipalities. Primary Health Care (PHC) units and emergency services in all cities are the gateways to TB cases. There is a reference hospital for TB cases

that require hospitalization related to severe drug reactions, extrapulmonary forms, patients co-infected with the human immunodeficiency virus (TB/HIV), and cases of resistance to the bacillus. In 2018, the total number of confirmed TB cases in the region, according to SINAN, was 537. The main outcomes were: 366 cures, 73 treatment dropouts, and 19 deaths from TB [22]. Directly Observed Treatment is performed in most municipalities, especially in cases of non-adherence to treatment or in cases of resistant TB.

### *11.3.3 Pre-implementation planning*

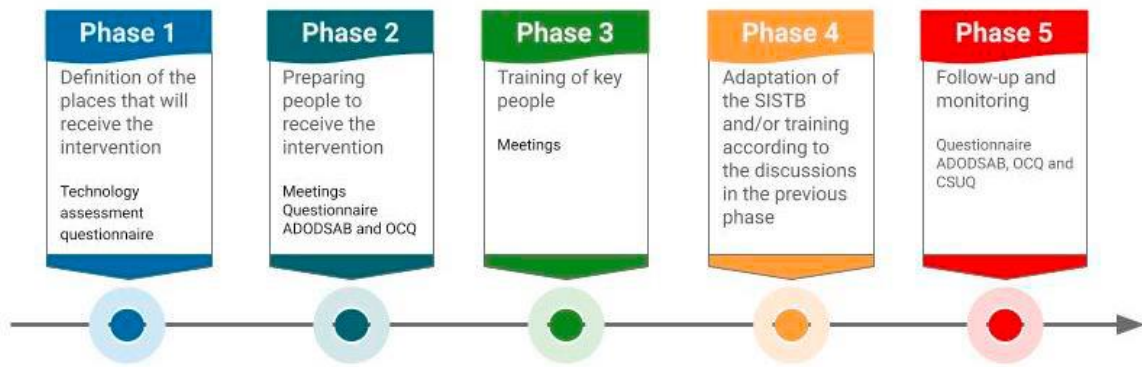
To fulfill the first objective, interviews will be carried out with managers and professionals of health services to characterize the way TB treatment is handled in the region. Due to the heterogeneity between the municipalities and the services offered by them, it is essential to know the particularities of each location. In this way, two types of interviews will be carried out, one to characterize how TB treatment is offered, how is the flow of intra-municipal and inter-municipal patients, and information flow. The other interview will characterize the services through which the patients' transit to evaluate the possible places that will receive the intervention. The interview will help to define the service in terms of the technological structure (e.g., technological infrastructure) to define the feasibility of implementing the intervention. Still in the pre-implementation phase, with these first data collected, it is intended to identify some possible barriers, so adjustments can be made, either in the SISTB or in the implementation protocol.

### *11.3.4 Implementation*

The implementation will be carried out in 5 phases: i) definition of the places that will receive the intervention; ii) preparing people to receive the intervention, holding a meeting to explain the purpose of the system and the implementation plan; iii) training of key people at each location; iv) adaptation of the SISTB and/or training according to the discussions in the previous phase; v) follow-up and monitoring.

Professionals who will have direct contact with SISTB will be identified and will respond to the Organizational Commitment Questionnaire (OCQ) [23] and the Questionnaire to evaluate the performance of health services in primary care in TB control applied to health professionals (ADODSAB) [24], as proposed by the protocol developed by Rui et al [25] to evaluate a software for monitoring the Directly Observed Treatment. The OCQ and ADODSAB questionnaires will be answered again, together with the Computer System

Usability Questionnaire (CSUQ) [26], after 3 months of using the system. Finally, results from each questionnaire will be compared with those of the moment before implementation. Fig. 1 illustrates the timeline of the implementation phases, when the questionnaires will be applied, and the meetings.



**Fig. 1.** Timeline of implementation phases

#### 11.3.5 Implementation analysis using RE-AIM

This study will be guided by the RE-AIM model to assess five dimensions (Reach, Effectiveness, Adoption, Implementation and Maintenance) [15]. RE-AIM was selected for this intervention because it addresses crucial points of implementation in real-world environments and identifies facilitating and hindering factors to achieve success, items relevant to external validity [27]. The analyzed items in each dimension were based on the RE-AIM [14, 15, 28] and on a literature review carried out by the authors to verify the use of the model to assess the implementation of SIS. Table 11 shows the study questions and their respective data sources.

**Table 11 - RE-AIM dimensions, study questions and data source**

Dimension	Question	Data source
Reach	Absolute number, proportion and characteristics of patients registered in SISTB	SISTB database
	Registration rate (were all patients undergoing treatment registered in SISTB?)	SISTB database / official record
Effectiveness	Has there been a change in organizational motivation?	Questionnaire OCQ
	Did the intervention affect service performance?	Questionnaire ADODSAB
	Abandonment rate (percentage of professionals who used the system in the first month and did not use it later)	SISTB database
	Overall satisfaction with SISTB	Questionnaire CSUQ
	SISTB usability	Questionnaire CSUQ

	Was the SISTB effective? Are modifications needed for greater effectiveness?	Meetings with staff and researchers
Adoption	Setting inclusion/exclusion criteria	Technology assessment questionnaire
	Number of eligible and invited settings	Technology assessment questionnaire
	Number of participating settings	SISTB database
	Characteristics of settings participating	Technology assessment questionnaire
	Staff inclusion/exclusion criteria	Research team follow-up / Meetings with staff and researchers
	Absolute number, proportion and characterization of staff who used the SISTB	Questionnaire ADODSAB / SISTB database
	Average of staff participating per setting	SISTB database
	System use	SISTB database
Implementation	Absolute number, frequency and duration of meetings for SISTB usage training	Research team follow-up
	Description of meetings to understand the implementation process	Research team follow-up
	Extent to which the protocol performed as expected	Research team follow-up
	What enabling factors were identified?	Research team follow-up / Meetings with staff and researchers
	What barriers to implementation were identified?	Research team follow-up / Meetings with staff and researchers
Maintenance	Continuity of the intervention after the research	SISTB database
	Reasons for discontinuity	Research team follow-up
	Program institutionalization	Research team follow-up
	Organizational abandonment rate	SISTB database
	What reinforcing factors were identified?	Research team follow-up / Meetings with staff and researchers

In the first dimension - scope -, it will be determined how many patients were registered in the system from the total number of patients undergoing treatment. The characterization of the patients treated who were registered will also be carried out. A measure of the delay in registration from diagnosis at the service will show how quickly the patient is entered into the system. By evaluating effectiveness, it is expected to know whether the intervention affected organizational motivation and service performance. Satisfaction with SISTB and the abandonment rate in the use of the system will also be evaluated. When evaluating the adoption, the exclusion criteria, the places included and the characterization of the participating professionals will be reported. The use of the system by the participants will be described with data collected from the SISTB database. Regarding the implementation, training sessions will be carried out with a group of professionals from each location. The professionals who participate in the meetings will be responsible for passing on the learning to co-workers. In the meetings, possible barriers or facilitators will also be investigated, whether in the implementation progress or related to the system. The last dimension, maintenance, will verify

if there was abandonment or continuity in the use of SISTB after training and the reasons for such decision. The reinforcing factors related to the use of the system will also be investigated

Data will be collected from several sources, including the SISTB database, the questionnaires and training meetings. Trained researchers will be responsible for collecting the data using electronic data capture (EDC) systems.

#### **11.4 Final considerations**

Information and communication technologies can contribute to the end of TB through improvements in the quality of patient care, surveillance, and service management [5]. This is a study to assess the implementation of a computerized system to track TB cases in a given region. The findings can help future systems implementation processes in environments with different levels of complexity.

Published evidence and best practices on digital health are essential aids to decision-makers and it is essential to disseminate the experience gained over time [5]. Implementation research comprises a research perspective that fosters collaboration between the actors involved, who must work together during research development to build trusting collaborations and encourage knowledge co-production [29].

Given the difficulty of integrating a new process or system within a health service, it must be considered that it is complex to characterize all the direct and indirect changes that the process will generate, especially when interventions interfere with people's routine. Interventions are not simple, and regardless of the restructuring that takes place, waiting for things to improve just by inserting the intervention is not enough. It is the responsibility of the implementation researcher to recognize and understand the repercussions that the execution of the new project can bring [13].

Finally, in this protocol, implementation research was used as a guide, as it seeks to work within the context of the real world, without trying to limit itself to the conditions of the population to avoid causal effects. Therefore, the population really represents the one that will be affected by the intervention [30]. The RE-AIM model was applied to assess the implementation process, as it is one of the most frequently applied [15], in addition to enabling a practical approach in the analysis of the resulting implications [13].

The forecast for the study conclusion is the end of 2023. As a result, we hope that the SISTB implementation will increase the positive outcomes of tuberculosis patients' treatment in the region of action. As a possible benefit for health professionals, it may help to organize

the data produced in the service routine and, for the patient, there may be an improvement in the assistance received and in treatment adherence. The results of this research may serve as a basis for other studies related to software implementation in health services.

### **11.5 References**

- [1] World Health Organization. (2021) Global Tuberculosis Report 2021. Global Tuberculosis Report. Geneva.
- [2] Organization, World Health. (1999) What is DOTS? What is DOTS? A Guide to Understanding the WHO-recommended TB Control Strategy Known as DOTS. Geneva.
- [3] World Health Organization. (2019) Global tuberculosis report 2019. Geneva.
- [4] World Health Organization. (2015) A global action framework for TB research in support of the third pillar of WHO's end TB strategy. Geneva.
- [5] World Health Organization, and European Respiratory Society. (2015) Digital health for the end TB strategy: an agenda for action. Geneva.
- [6] World Health Organization. (2017) Digital health for the end TB strategy: progress since 2015 and future perspectives: meeting report, 7 -8 February 2017. Geneva.
- [7] Crepaldi, Nathalia Yukie, Vinicius Costa Lima, Filipe Andrade Bernardi, Luiz Ricardo Albano Santos, Verena Hokino Yamaguti, Felipe Carvalho Pellison, Tiago Lara Michelin Sanches, et al. (2019) SISTB: an ecosystem for monitoring TB. *Procedia Computer Science* 164: 587–594. <https://doi.org/10.1016/J.PROCS.2019.12.224>.
- [8] Crepaldi, Nathalia Yukie, Inácia Bezerra de Lima, Fernanda Bergamini Vicentine, Lídia Maria Lourenço n Rodrigues, Verena Hokino Yamaguti, Tiago Lara Michelin Sanches, Antonio Ruffino -Netto, Domingos Alves, and Rui Pedro Charters Lopes Rijo. (2017) Satisfaction evaluation of health professionals in the usability of software for monitoring the tuberculosis treatment. *Procedia Computer Science* 121: 889–896. <https://doi.org/10.1016/j.procs.2017.11.115>.
- [9] Crepaldi, Nathalia Yukie, Inacia Bezerra de Lima, Fernanda Berga mini Vicentine, Lídia Maria Lourençon Rodrigues, Tiago Lara Michelin Sanches, Antonio Ruffino-Netto, Domingos Alves, and Rui Pedro Charters Lopes Rijo. (2018) Towards a Clinical Trial Protocol to Evaluate Health Information Systems: Evaluation of a Computerized System for Monitoring Tuberculosis from a Patient Perspective in Brazil. *Journal of Medical Systems* 42: 8. <https://doi.org/10.1007/s10916-018-0968-8>.
- [10] World Health Organization. (2012) Electronic recording and reporting for tuberculosis care and control. Geneva: World Health Organization.

- [11] World Health Organization. (2020) New research tool supports scale-up of digital technologies to End TB. Geneva.
- [12] World Health Organization & UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases. (2011) Implementation research for the control of infectious diseases of poverty: strengthening the evidence base for the access and delivery of new and improved tools, strategies and interventions. Geneva.
- [13] Peters, David, Nhan Tran, Taghreed Adam, Alliance for Health Policy and Systems Research, and World Health Organization. (2013) Implementation research in health: a practical guide. Geneva.
- [14] Bakken, Suzanne, and Cornelia M. Ruland. (2009) Translating Clinical Informatics Interventions into Routine Clinical Care: How Can the RE-AIM Framework Help? *Journal of the American Medical Informatics Association* 16: 889–897. <https://doi.org/10.1197/JAMIA.M3085>.
- [15] Glasgow, Russell E., Samantha M. Harden, Bridget Gaglio, Bor sika Rabin, Matthew Lee Smith, Gwendolyn C. Porter, Marcia G. Ory, and Paul A. Estabrooks. (2019) RE-AIM planning and evaluation framework: Adapting to new science and practice with a 20 -year review. *Frontiers in Public Health* 7: 64. <https://doi.org/10.3389/fpubh.2019.00064>.
- [16] Pellison, Felipe Carvalho, Rui Pedro Charters Lopes Rijo, Vinicius Costa Lima, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Rafael Mello Galliez, Afrânio Kritski, Kumar Abhishek, and Domingos Alves. (2020) Data Integration in the Brazilian Public Health System for Tuberculosis: Use of the Semantic Web to Establish Interoperability. *JMIR Medical Informatics* 8. <https://doi.org/10.2196/17176>.
- [17] Lima, Vinicius Costa, Domingos Alves, Felipe Carvalho Pellison, Vinicius Tohoru Yos hiura, Nathalia Yukie Crepaldi, and Rui Pedro Chartes Lopes Rijo. (2018) Establishment of Access Levels for Health Sensitive Data Exchange through Semantic Web. *Procedia Computer Science* 138: 191–196. <https://doi.org/10.1016/J.PROCS.2018.10.027>.
- [18] Lima, Vinicius Costa, Felipe Carvalho Pellison, Filipe Andrade Bernardi, Domingos Alves, and Rui Pedro Charters Lopes Rijo. (2021) Security Framework for Tuberculosis Health Data Interoperability Through the Semantic Web. *International Journal of Web Portals* 13: 36–57. <https://doi.org/10.4018/IJWP.2021070103>.
- [19] Lima, Vinícius, Felipe Pellison, Filipe Bernardi, Isabelle Carvalho, Rui Rijo, and Domingos Alves. (2019) Proposal of an integrated decision support system for Tuberculosis based on Semantic Web. *Procedia Computer Science* 164: 552–558. <https://doi.org/10.1016/J.PROCS.2019.12.219>.



- [20] Hokino Yamaguti, Verena, Domingos Alves, Rui Pedro Charters Lopes Rijo, Newton Shydeo Brandão Miyoshi, and Antônio Ruffino -Netto. (2020) Development of CART model for prediction of tuberculosis treatment loss to follow up in the state of São Paulo, Brazil: A case– control study. *International Journal of Medical Informatics* 141. <https://doi.org/10.1016/J.IJMEDINF.2020.104198>.
- [21] Instituto Brasileiro de Geografia e Estatística. (2010) Censo demográfico 2010.
- [22] Brazil. (2021) Casos de tuberculose - Sistema de Informação de Agravos de Notificação - Sinan Net.
- [23] Mowday, Richard T., Richard M. Steers, and Lyman W. Porter. (1979) The measurement of organizational commitment. *Journal of Vocational Behavior* 14: 224–247. [https://doi.org/10.1016/0001-8791\(79\)90072-1](https://doi.org/10.1016/0001-8791(79)90072-1).
- [24] Villa, Tereza Cristina Scatena, and Antônio Ruffino-Netto. (2009) Performance assessment questionnaire regarding TB control for use in primary health care clinics in Brazil. *Jornal Brasileiro de Pneumologia* 35: 610–612. <https://doi.org/10.1590/S1806-37132009000600014>.
- [25] Rijo, Rui Pedro Charters Lopes, Nathalia Yukie Crepaldi, Fernanda Bergamini, Lídia Maria, Lourençon Rodrigues, Inácia Bezerra De Lima, Gleici Silva Castro Perdoná, and Domingos Alves. (2017) Impact assessment on patients’ satisfaction and healthcare professionals’ commitment of software supporting Directly Observed Treatment, Short - course: A protocol proposal. *Health Informatics Journal*. <https://doi.org/10.1177/1460458217712057>.
- [26] Lewis, James R. (2009) IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction* 7: 57–78. <https://doi.org/10.1080/10447319509526110>.
- [27] Tuot, Delphine S., Clare Liddy, Varsha G. Vimalananda, Jennifer Pecina, Elizabeth J. Murphy, Erin Keely, Steven R. Simon, Frederick North, Jay D. Orlander, and Alice Hm Chen. (2018) Evaluating diverse electronic consultation programs with a common framework. *BMC Health Services Research* 18: 814. <https://doi.org/10.1186/s12913-018-3626-4>.
- [28] Brito, Fabiana Almeida, Tânia Rosane Bertoldo Benedetti, Camila Tomicki, Lisandra Maria Konrad, Paula Fabrício Sandresch i, Sofia Wolker Manta, and Fabio Almeida. (2018) Translation and adaptation of the RE-AIM Check List for Brazilian reality. *Revista Brasileira de Atividade Física & Saúde* 23: 1–8. <https://doi.org/10.12820/rbafs.23e0033>.
- [29] Theobald, Sally, Neal Brandes, Margaret Gyapong, Sameh El -Saharty, Enola Proctor, Theresa Diaz, Samuel Wanji, et al. (2018) Implementation research: new imperatives and

opportunities in global health. *The Lancet* 392: 2214–2228. [https://doi.org/10.1016/S0140-6736\(18\)32205-0](https://doi.org/10.1016/S0140-6736(18)32205-0).

[30] Peters, David H., Taghreed Adam, Olakunle Alonge, Irene Akua Agyepong, and Nhan Tran. (2013) Implementation research: what it is and how to do it. *British Medical Journal* 347:731–736. <https://doi.org/10.1136/BMJ.F6753>.

## 12 OTHER PRODUCTIONS

The data management for this doctoral dissertation emphasizes free access to the proposed contents, aimed at enhancing the literature related to the studied problem and enabling the application of the suggested solution methods in practical projects in other areas with an implementation environment akin to that investigated in the research. Within the scope of the doctoral student's engagement and involvement in his managerial functions, several adjacent productions were carried out that are based on or complement the results presented in the previous chapters. The main generated data were classified into two groups distributed in six distinct categories described in the following subsections.

### 12.1 Sharing and Implementation Resources

This category covers all materials and tools used in the process of sharing and implementing research outputs. That includes:

- Data management plans, metadata, analysis and scientific communication materials and training
- Source codes of methods and models, mobile applications, and analyzes developed in the presented application scenarios.
- Data sets, processes, programming instances used to compare methods, in pdf, .txt, .rdf, .owl, or .csv format.

All of the above products and their respective information are available in the following sharing data platform repositories:

- **Digital Health Atlas:** *“a WHO global technology registry platform aiming to strengthen the value and impact of digital health investments, improve coordination, and facilitate institutionalization and scale.”*
- **FAIRSharing:** is a comprehensive catalog that details and connects data standards, databases, and data policies in the scientific fields, helping users identify and utilize these resources efficiently.
- **Github:** A widely-used web-based platform for version control and collaborative software development. Its use in this research project has enabled transparent, organized, and collaborative development of digital health solutions, contributing significantly to the reproducibility and accessibility of research outputs.

- **LattesData:** is a Brazilian open access repository for scientific research data, focusing on preservation and sharing of data from CNPq-funded projects to support Open Science and enhance research reproducibility and efficiency.
- **Instituto Nacional da Propriedade Industrial (INPI):** The National Institute of Industrial Property is the Brazilian agency responsible for managing the country's system of granting and ensuring intellectual property rights. This includes the registration of trademarks, patents, industrial designs, geographical indications, and copyright protection related to software and topographies of integrated circuits.

Table 12 details the distribution of the scientific production generated throughout this doctoral dissertation. This table aims to provide a clear and structured overview of the various categories of data and resources generated and shared. We also highlight the location of these resources across the mentioned data-sharing platforms, reflecting a commitment to open access and the promotion of open science.

**Table 12 - Doctoral Dissertation Sharing and Implementation Resources**

Platform/Repository	Product	Access link	Related Doctoral Thesis Chapter's
Digital Health Atlas	Brazilian Rare Disease Portal	<a href="https://digitalhealthatlas.org/p/BRej2XRej">https://digitalhealthatlas.org/p/BRej2XRej</a>	IV, VI and VII
FAIRSharing	Technological Infrastructure Brazilian Rare Disease Network	<a href="https://doi.org/10.25504/FAIRsharing.d7b6c8">https://doi.org/10.25504/FAIRsharing.d7b6c8</a>	VI and VII
Github	TB and RD application programming codes	<a href="https://github.com/filipepaulista12">https://github.com/filipepaulista12</a>	VIII and IX
LattesData	National Rare Disease Registry	<a href="https://doi.org/10.57810/lattesdata/XEL53O">https://doi.org/10.57810/lattesdata/XEL53O</a>	V, X
INPI	Brazilian Rare Disease Portal	Appendix A - Registration number: BR5120220032899	VII
	SISTB	Appendix B - Registration number: BR512023002666 2	XI

Source: The Author

Each table row represents a vital component of the research conducted and a significant contribution to the field of digital health, facilitating the application of proposed methods in practical projects in other areas and encouraging ongoing research and innovation. The

distribution of these resources is critical to understanding the breadth and depth of the work conducted and providing a valuable resource for researchers, practitioners, and students interested in delving deeper into the results and methodologies developed in this dissertation.

The sets of instances, codes, figures, tables and results do not cover any ethical or legal issues. Therefore, they are suitable for use in research projects, as long as the source (doctoral thesis or articles from each chapter of the thesis) is properly cited. They can also be replicated in public documents, such as scientific articles and technical reports, as long as the source is cited.

## **12.2 Communications and Dissemination Resources**

This category is vital for validating and disseminating the products of our work. It provides tangible evidence of research results and serves as a means to communicate the findings to the academic world and potential practical applicators. Publications are essential for establishing the credibility and impact of your work and advancing knowledge in digital health. The derivated scientific products related to this doctoral thesis includes:

- **Articles Outside the Thesis Theme:** These include work that, although related, is not directly aligned with the central theme of the thesis. They are presented in Table 13 and address peripheral topics, explore alternative applications of the technologies or methods studied, and investigate related theoretical or methodological issues. It reflects the interdisciplinary impact of our research and underscores the versatility of our findings. They explore alternative applications of the technologies and methodologies studied, delving into peripheral topics and investigating related theoretical or methodological issues. The breadth of our research has stimulated discussions in various domains, showcasing the wide-reaching implications of our work.
- **Mentorship and Academic Development:** This category reflects the significant contributions to student academic growth through the supervision of undergraduate theses, guidance in scientific initiation projects, and mentorship of master's students. These mentorship activities, described in Table 14, are pivotal for students' academic and professional advancement and enriches the research work, bringing new perspectives and opportunities for collaboration in digital health. It empowered students with practical skills, theoretical insights, and research experience, laying the broader impact and applicability of the thesis results.
- **Conference Papers:** These are papers that have been presented at academic conferences. They were essential for disseminating research results and receiving feedback from the

scientific community. Conferences often allow discussion of preliminary or ongoing results and receiving valuable insights from other researchers. This active and reciprocal participation, presented in Table 15, supported the model improvement and also contributed to its refinement, strengthening its applicability and impact within the academic discourse on digital health research.

**Table 13 - Mentorship and Academic Development Resources**

<b>Activity/ Role/Tool</b>	<b>Title</b>	<b>Description</b>	<b>Access link</b>
Project management and Collaboration - Trello	Documentation about techniques, business intelligence, benchmarking, process and monitoring usage	Trello played a crucial role in research efficiency through utilizing digital tools for managing research activities, collaboration, and tracking progress. Number of Collaborators: 20	RD: <a href="https://trello.com/w/rarasrp">https://trello.com/w/rarasrp</a>  TB: <a href="https://trello.com/w/tblis">https://trello.com/w/tblis</a>
Orientation and Co-orientation - Undergraduate Thesis	ELISIOS - Multiplatform app for calculating scores applied to predict the risk of abandonment of patients with tuberculosis Leonardo Garmatz Berres (2019)	As the advisor of the ELISIOS app, we demonstrate an initial integration of innovative technological tools into health data infrastructure to support clinical decisions, like database creation and data integration. As part of the broader ELISIOS-TB study involving multiple Brazilian institutions, this collaborative effort reflects the multi-disciplinary approach I advocate in this thesis, highlighting the emphasis on infrastructure and data management in healthcare	<a href="https://drive.google.com/drive/folders/1iRGKAivwYBL2z8ixGt1PVJSX32gdLoP?usp=sharing">https://drive.google.com/drive/folders/1iRGKAivwYBL2z8ixGt1PVJSX32gdLoP?usp=sharing</a>
	TBI Score - Mobile application for calculating scores to help diagnose tuberculosis in children Danilo Sampaio (2021)	This project focused on developing a mobile application based on the scoring system outlined in the 2019 Brazilian Tuberculosis Control Guidelines. The application is designed to assist in diagnosing pulmonary tuberculosis in children and adolescents, incorporating clinical, radiological, and epidemiological data aspects. The alignment of the app with the structured guidelines from the Ministry of Health demonstrates the practical application of national health guidelines in digital form, enhancing the accessibility and usability of these guidelines for healthcare professionals.	
To be continued			

Conclusion			
<b>Activity/ Role/Tool</b>	<b>Title</b>	<b>Description</b>	<b>Access link</b>

<p>From raw data to FAIR data: the FAIRification workflow for tuberculosis research Gabriel Sartoretto (2023)</p>	<p>This project evaluated Brazilian TB data within health information systems against FAIR standards, uncovering issues like the absence of unique identifiers and a prevalence of non-intuitive and inconsistent data. My guidance aimed at enhancing data quality and promoting the adoption of FAIR principles in health data systems aligns with my thesis's emphasis on improving the management and usability of health information. The project's findings and strategies contribute to strengthening national health information systems by highlighting current limitations and suggesting improvements.</p>	
<p>Tuberculosis data quality analysis: Case study of a center Brazilian clinical research. Gabriel Modina (2023)</p>	<p>This project emphasized the heterogeneity and complexity of clinical TB research data, challenging its interoperability and quality. The project reflected my thesis's goal of enhancing health data management and usage under FAIR guidelines by evaluating data quality attributes and identifying patterns that don't meet established standards. This work contributed to understanding data quality mechanisms in TB services, resonating with a focus on improving decision-making and research outcomes through rigorous data quality assessment in healthcare research networks.</p>	
<p>Digital Strategies for the Rare Disease Information Portal: Expanding Reach and Impact Vitor Berardi (2023)</p>	<p>This study focused on developing digital strategies to enhance the visibility and impact of the RARAS portal for RD in Brazil. Using data analysis tools and digital marketing strategies to improve the portal's reach and effectiveness exemplifies the aspects of the RE-AIM framework. Overall, this project exemplified the critical role of digital strategies in health communication and the importance of maintaining high data quality standards.</p>	

Source: The Author



**Table 14** - Articles outside the theme thesis

Journal	Publication/Product	Collaboration Description	Access link
Nature Scientific Reports - Journal (ISSN: 2045-2322)	REDbox: a comprehensive semantic framework for data collection and management in tuberculosis research (LIMA et al., 2021b)	This study, focusing on developing a semantic framework for TB research, aligns closely with my thesis's emphasis on enhancing data management and governance in digital health research. My participation in the REDbox project involved creating a framework that addresses critical aspects of data collection and management, a core focus of my doctoral work. The semantic framework developed in REDbox aimed to streamline and improve TB research's data handling quality. It also exemplified the practical implementation of the theoretical concepts I explored in my thesis, thereby contributing to the broader field of digital health informatics and emphasizing the importance of robust data management systems in enhancing the efficacy of health research.	<a href="https://doi.org/10.1038/s41598-023-33492-6">https://doi.org/10.1038/s41598-023-33492-6</a>
Revista de Saúde Digital e Tecnologias Educacionais - RESDITE - (ISSN: 2525-9563)	The Collaborative Data Gathering And Audit Process In The Covid-19 Brazil Portal: Health Information In The Pandemic (CARVALHO et al., 2022a)	In this collaborative work, my role as IT coordinator significantly intersects with the themes explored in my doctoral thesis, which focused on data gathering and auditing processes in the context of the COVID-19 pandemic. The study emphasizes that the decentralization and lack of standardization in information reporting pose significant challenges, particularly in automating data collection processes. The collaborative effort succeeded in maintaining daily analyses of municipal data, ensuring the dissemination of reliable and high-quality information to the public. My involvement in this project entailed contributing to developing and refining methodologies for data collection and auditing, which are central aspects of my doctoral research. This work not only addresses the immediate challenge of pandemic response but also contributes valuable insights into the management of health information during public health crises.	<a href="http://periodicos.ufc.br/resdite/article/view/62767/226150">http://periodicos.ufc.br/resdite/article/view/62767/226150</a>
To be continued			

Continuation			
Journal	Publication/Product	Collaboration Description	Access link
Orphanet Journal of Rare Diseases - (ISSN: 1750-1172)	Epidemiology of rare diseases in Brazil: protocol of the Brazilian Rare Diseases Network (RARAS-BRDN) (FÉLIX et al., 2022)	My collaboration with the RARAS networks in this work was closely connected to the central themes of my doctoral thesis, focusing on the epidemiology of RD in Brazil from the MDS perspective. It contributed to developing the protocol for the RARAS. It aligns with the emphasis on epidemiological and offers a practical application of the theoretical concepts I explored in my thesis, particularly regarding data collection, quality, and management. Furthermore, the collaborative nature of the RARAS project reflected my thesis's focus on collaborative and participatory approaches in health research.	<a href="https://doi.org/10.1186/s13023-022-02254-4">https://doi.org/10.1186/s13023-022-02254-4</a>
Revista Ibérica de Sistemas e Tecnologias de Informação - (ISSN-e: 1646-9895)	More than words: an analysis of Brazilian emotions during COVID-19 (BERNARDI et al., 2021)	My leadership in this work dovetails with the key themes that focus on leveraging digital data for health research. This study's focus on analyzing the emotions of Brazilians during the COVID-19 pandemic complements my thesis's broader exploration of digital health and public health informatics. The study's use of information systems to capture and interpret public sentiment during a health crisis showcases the importance of digital tools in understanding and managing public health phenomena. It emphasizes the critical need for reliable and accessible information, the importance of transparent information systems, and the empowerment of healthcare professionals and local administrators in making informed decisions across various health sectors. This approach underscores the significance of effective information management and public sentiment analysis in health crisis situations.	<a href="https://repositorio.usp.br/item/003075794">https://repositorio.usp.br/item/003075794</a>
			To be continued

Continuation

Journal	Publication/Product	Collaboration Description	Access link
Journal of Multiprofessional Health Research - (ISSN: 2675-8849)	The relevance of digital health strategies to support tuberculosis services in the COVID-19 pandemic context (Letter) (LIMA et al., 2021c)	This letter explores the critical intersection of TB management and digital health strategies during the COVID-19 pandemic. My contribution to this project involved discussing various innovative approaches implemented by the LIS. These include a platform for remote monitoring of TB patients, developing an interoperability layer for health information systems, predictive models for treatment adherence, and an alert system for COVID-19-related changes.	<a href="https://repositorio.usp.br/item/003073667">https://repositorio.usp.br/item/003073667</a>
Methods of Information in Medicine - (ISSN: 0026-1270)	A Permissioned Blockchain Network for Security and Sharing of De-identified Tuberculosis Research Data in Brazil (LIMA et al., 2020)	Using blockchain technology, this study aims to enhance the security and sharing of TB research data in Brazil. This effort complements my doctoral research, which explores using advanced technologies to improve data management and security in health research. The project's emphasis is on ensuring data transparency, accountability, availability, and integrity. My role in this project was to develop a blockchain-based approach for creating a secure, decentralized, and de-identified dataset for TB research described in the before-published protocol (BERNARDI et al., 2019). This aligns with my thesis's focus on digital health governance, data security, and innovative technological solutions in healthcare research.	<a href="https://doi.org/10.1055/s-0041-1727194">https://doi.org/10.1055/s-0041-1727194</a>
International Journal of Web Portals - (ISSN: 1938-0194)	Security Framework for Tuberculosis Health Data Interoperability Through the Semantic Web (LIMA et al., 2021d)	This study addresses the vital challenge of interoperability in health information systems, focusing on TB data exchange. The study's emphasis on leveraging Semantic Web technologies aligns with the thesis's objective of improving data sharing capabilities, data quality, and completeness while addressing the critical concern of data security. It highlights the significance of creating trust in health data and promoting its use in various contexts, especially where data dissemination is challenging without significant technological support.	<a href="https://doi.org/10.4018/IJWP.2021070103">https://doi.org/10.4018/IJWP.2021070103</a>
To be continued			

Conclusion			
Journal	Publication/Product	Collaboration Description	Access link

<p>The Journal of the Brazilian Society of Tropical Medicine - (ISSN: 1678-9849)</p>	<p>Brazil: the emerging epicenter of COVID-19 pandemic (NEIVA et al., 2020)</p>	<p>This research focuses on analyzing the progression of the COVID-19 pandemic in Brazil and the implications of policy decisions on public health. My role in this project involved contributing to the data analysis and statistical evaluation of COVID-19's spread in various countries, including a detailed exploration of the pandemic's impact within Brazilian states. This complements my doctoral research, which delves into the utilization of digital tools and data-driven approaches for managing public health crises.</p>	<p><a href="https://doi.org/10.1590/0037-8682-0550-2020">https://doi.org/10.1590/0037-8682-0550-2020</a></p>
<p>JMIR Medical Informatics - (ISSN: 2291-9694)</p>	<p>Data Integration in the Brazilian Public Health System for Tuberculosis: Use of the Semantic Web to Establish Interoperability (PELLISON et al., 2020)</p>	<p>This research addressing the interoperability challenges in health information systems, particularly in the context TB, complements my thesis's focus on enhancing data management and governance in digital health according to the patterns definition. My role in this project entailed evaluating an interoperability solution for TB treatment and follow-up using Semantic Web technology patterns. The methodological approach involving Basic Formal Ontology to tag Brazilian TB applications and develop an interoperability layer reflects my interest in employing ontology-based solutions for health data integration.</p>	<p><a href="https://doi.org/10.2196/17176">https://doi.org/10.2196/17176</a></p>

Source: The Author

Table 15 - Conference Papers Collaboration

Journal Collaboration	Publication/Product	Description	Access link
International Conference on Health and Social Care Information Systems and Technologies - HCIST	a) Quality analysis and study of tuberculosis diagnostic data MIOTO et al., 2024)	In <b>2023</b> , the HCIST conference papers closely aligned with the author's doctoral thesis themes. The author's roles in Projects A and B reflected the thesis's focus on health data quality, digital infrastructure, and TB diagnosis support, contributing through mentorship, ideation, and validation. The collaboration in Project C also aligned with the thesis, exploring digital health solutions for RD research. These activities underscored the author's integrated approach to health informatics, effectively bridging academic research with practical healthcare applications.	In Press
	b) OUTB: application for decision-support in the outcomes of Tuberculosis (MOZINI et al., 2024)		
	c) ICD-10 - ORPHA: An Interactive Complex Network Model for Brazilian Rare Diseases (NEIVA et al., 2024)		
	d) Informed Consent Form Automated Validation, The Brazilian Rare Disease Network Case Proposal. (NEIVA et al., 2023)	In <b>2022</b> , the author's collaborative involvement in projects D and E, presented at the HCIST conference, significantly aligned with the themes of their doctoral thesis. Project D directly connects to the thesis's focus on enhancing ethical data governance and informed consent practices in digital health research, particularly RD. In Project E, the author's role in validating the research protocol complemented the thesis's exploration of data infrastructure and digital health practices across various regional landscapes in Brazil.	<a href="https://doi.org/10.1016/j.procs.2023.01.445">https://doi.org/10.1016/j.procs.2023.01.445</a>
	e) Scaling laws and spatial effects of Brazilian health regions: a research protocol. (SOARES et al., 2023)		<a href="https://doi.org/10.1016/j.procs.2023.01.417">https://doi.org/10.1016/j.procs.2023.01.417</a>
			To be continued

Continuation

Journal Collaboration	Publication/Product	Description	Access link
	f) TBI Score - use of a mobile score system to aid the diagnosis of tuberculosis in children in Brazil. (BERNARDI et al.,2022)	<p>In <b>2021</b>, the author's involvement in projects F, I, and J at the HCIST conference showcased their expertise in mentorship developing digital healthcare solutions, particularly for TB care. Project F's development of mobile tools for TB diagnosis in children and Project I's mobile application for TB treatment adherence highlighted the practical application of digital technology in healthcare. Project J further emphasized this theme by creating integrated data systems for TB care. Projects G and K also focused on enhancing healthcare research infrastructure and disseminating health information effectively. Project G advanced data management for RD research in Brazil, while Project K contributed to public health communication through a web portal for COVID-19 information in Brazil. Project H complemented these themes by exploring data analytics in tracking the COVID-19 pandemic. These projects collectively highlighted the author's comprehensive contributions to digital health research.</p>	<a href="https://doi.org/10.1016/j.procs.2021.12.041">https://doi.org/10.1016/j.procs.2021.12.041</a>
	g) The minimum dataset for rare diseases in Brazil: a systematic review protocol. (BERNARDI et al.,2022)		<a href="https://doi.org/10.1016/j.procs.2021.12.034">https://doi.org/10.1016/j.procs.2021.12.034</a>
	h) Unsupervised analysis of COVID-19 pandemic evolution in brazilian states. (CASSÃO et al., 2022)		<a href="https://doi.org/10.1016/j.procs.2021.12.061">https://doi.org/10.1016/j.procs.2021.12.061</a>
	i) My Latent Tuberculosis Treatment - mobile application to assist in adherence to latent tuberculosis treatment (ARNIZANT et al., 2022)		<a href="https://doi.org/10.1016/j.procs.2021.12.059">https://doi.org/10.1016/j.procs.2021.12.059</a>
	j) A computational infrastructure for semantic data integration towards a patient-centered database for Tuberculosis care (LIMA et al., 2022)		<a href="https://doi.org/10.1016/j.procs.2021.12.033">https://doi.org/10.1016/j.procs.2021.12.033</a>
			To be continued

Continuation			
Journal Collaboration	Publication/Product	Description	Access link

	k) COVID-19 BR: A web portal for COVID-19 information in Brazil (CARVALHO et al., 2022b)		<a href="https://doi.org/10.1016/j.procs.2021.12.045">https://doi.org/10.1016/j.procs.2021.12.045</a>
	l) Operational modeling for testing diagnostic tools impact on tuberculosis diagnostic cascade: A model design (DA COSTA et al., 2021)	In <b>2020</b> , the author's collaboration on Project L at the HCIST conference focused on a literature review aimed at improving TB care diagnostics and interventions using innovative technology. This project emphasized the identification of needs and barriers in accessing TB data in Brazil, laying the groundwork for validating a proposed model in future research.	<a href="https://doi.org/10.1016/j.procs.2021.01.214">https://doi.org/10.1016/j.procs.2021.01.214</a>
International Conference on Informatics, Management and Technology in Healthcare - ICIMTH	m) Development of a Mobile Application with Health Guidelines for TB Patients Care (REIS et al., 2023)	In <b>2023</b> , at the ICIMTH, the author collaborated on project M. The collaboration in this project showcased the correlation of the author's thesis results to enhancing patient engagement and care in TB treatment through the use of innovative mobile technology aligned with updated official guidelines. This project results from the ongoing focus on digital health solutions, which involves creating another mobile application designed to provide comprehensive health guidelines for patients undergoing TB treatment.	<a href="https://doi.org/10.3233/shti230509">https://doi.org/10.3233/shti230509</a>
	n) Blockchain based Network for Tuberculosis: Data Sharing Initiative in Brazil. Health Informatics (BERNARDI et al., 2019)	In <b>2019</b> , at the ICIMTH, the author was the key ideator of the project N. This initiative was the first driving force that reflects the author's innovative approach to health informatics in their thesis. Focused on developing a blockchain-based network to facilitate secure and efficient data sharing in Brazil's TB research context, this result highlights commitment to employing cutting-edge technology solutions to address the challenges of data management and sharing in healthcare.	<a href="https://doi.org/10.3233/shti210099">https://doi.org/10.3233/shti210099</a>
To be continued			

Conclusion			
Journal Collaboration	Publication/Product	Description	Access link
Conference on Health	p) A Mechanism for Verifying the Integrity and Immutability of Tuberculosis Data	At the dHealth in <b>2021</b> , the author collaborated on project P testing and validating previous hypotheses related to data integrity in the healthcare	<a href="https://doi.org/10.3233/shti190069">https://doi.org/10.3233/shti190069</a>

Informatics meets Digital Health - dHealth	Using IOTA Distributed Ledger Technology (LIMA et al., 2021e)	domain, specifically focusing on TB data. The project utilized distributed ledger technology to ensure the integrity and immutability of TB data, showcasing the author's engagement in applying technological solutions to enhance one of the dimensions of data quality on the security in health informatics.	
International Conference on Computational Science - ICCS	q) A Web Portal for Real-Time Data Quality Analysis on the Brazilian Tuberculosis Research Network: A Case Study (CASSÃO et al., 2023)	In ICCS 2023, the author focused on developing a web portal for real-time data quality analysis within the Brazilian Tuberculosis Research Network project (Q) providing a valuable infrastructure for enhancing research data monitoring. Also, project R explored supervised machine-learning techniques for data features discovering rare autoimmune diseases. This presented a promising approach to support the challenging diagnosis with real data.	<a href="https://doi.org/10.1007/978-3-031-36024-4_24">https://doi.org/10.1007/978-3-031-36024-4_24</a>
	r) Supervised Machine Learning Techniques Applied to Medical Records Toward the Diagnosis of Rare Autoimmune Diseases (ANDRADE MARTINS et al., 2023)		<a href="https://doi.org/10.1007/978-3-031-36024-4_13">https://doi.org/10.1007/978-3-031-36024-4_13</a>
	s) Knowledge Discovery in Databases: Comorbidities in Tuberculosis Cases (CARVALHO et al., 2022c)	In ICCS 2022, the authors showcased the evolution of TB database analysis and contributed to the broader understanding of condition health implications. Similarly, Project T, communicated the progress of the computational infrastructure for the RARAS networks in Brazil from validated structures focused on ensuring data privacy and protection from participating health facilities.	<a href="https://doi.org/10.1007/978-3-031-08757-8_1">https://doi.org/10.1007/978-3-031-08757-8_1</a>
	t) National network for rare diseases in Brazil: the computational infrastructure and preliminary results (YAMADA et al., 2022)		<a href="https://doi.org/10.1007/978-3-031-08757-8_4">https://doi.org/10.1007/978-3-031-08757-8_4</a>

Source: The Author



## 13 CONCLUSIONS AND FUTURE RESEARCH

### 13.1 Conclusions

In this concluding chapter, we address the significant contributions of the research, particularly in filling critical gaps identified in the academic literature and practical applications, with a specific focus on the Brazilian health systems. The dissertation significantly enhances the understanding of data quality and governance within the field of health research, especially concerning TB and RD. Through a comprehensive review and the introduction of novel frameworks, we bridge crucial gaps in digital health governance, a matter of particular relevance given the unique challenges posed by the socio-economic and geographical diversity of Brazil. The research underscores the imperative for standardized practices and enhanced collaborative efforts to augment data quality in Brazilian health research.

Despite the focus on health research, the presentation of practical implications is extensive and notably pertinent to health policy, clinical practice, and the design of health information systems. The proposed models, including the FAIRification workflow and a comprehensive governance framework, directly apply to enhancing health data management and sharing in Brazil. These models provide health practitioners and policymakers with robust frameworks to improve the accuracy, accessibility, and utility of health data, thereby leading to better clinical decision-making and policy development. For system developers, our findings offers clear guidelines for creating efficient and user-friendly health information systems.

Methodologically, the dissertation represents a significant advancement in health informatics research by adopting a mixed-methods approach that seamlessly integrates qualitative and quantitative data. This methodological integration allows for a nuanced understanding of the complexities inherent in digital health, facilitating an in-depth exploration of issues such as data quality and system interoperability. It emphasizes the critical importance of versatile research approaches in health informatics, particularly when dealing with diverse and complex health data sets.

This integrative strategy is particularly effective in the contemporary health research landscape, where the intricate interplay between digital data and public health demands in-depth analysis and a holistic understanding. Data is not just a collection of numbers or facts; it is deeply interconnected with public health outcomes and decisions. Therefore, understanding this data requires an in-depth, detailed analysis (to uncover specific patterns and trends) and a holistic view (to understand the broader impacts and implications for public health). The

"integrative strategy" combines various methods and perspectives to achieve this dual understanding, ensuring that digital health initiatives are data-driven and contextually informed.

The research also navigates through challenges like integrating diverse data sources and maintaining patient confidentiality while seizing opportunities to improve public health outcomes through digital innovation. Robust data governance is crucial, ensuring the integrity and quality of the data collected. In addition to the robust data governance, the study's key elements include innovative data analysis methods, mobile applications, and addressing the challenges and opportunities in the digital health implementation to guide the planning and evaluation programs of the health research framework.

By embracing this comprehensive framework, the research endeavor yields valuable insights and practical strategies to enhance the efficiency of research processes in digital health initiatives. A pivotal component of this framework involves harnessing the power of real-world data to inform the selection of research sites, thereby optimizing the relevance and impact of research endeavors. This is achieved through the creation of research repositories curated to enable the identification of patient cohorts for prospective studies, providing a dependable indicator of their prospective success.

Furthermore, within the network infrastructure, security takes precedence, with stringent measures in place, including data obfuscation, to safeguard against the inadvertent reidentification of patients and to uphold data integrity. Additionally, the research underscores the importance of data quality and semantic mapping. Data from diverse institutions undergo a harmonization process, mapped to agreed-upon terminologies, facilitating seamless queries within a federated network.

Table 16 provides a concise summary of the main contributions and recommendations derived from each research. Each row corresponds to a specific objective or hypothesis explored in the study, outlining the key findings and suggested actions. The "Main Contributions" column encapsulates the significant outcomes and advancements achieved in each chapter, highlighting the core impact of the research. Complementing this, the "Recommendations" column offers practical insights and guidance based on the study's outcomes, providing a roadmap for future endeavors in the field. Additionally, the "Related Chapter/Publication" column references the specific chapters or publications where detailed information on each objective or hypothesis can be found, facilitating further exploration and validation of the presented insights.

Table 16 - Summary of main contributions and recommendations

Objective/Hypothesis	Main Contributions	Recommendations	Related Chapter/ Publication
A - Conduct a literature review on data quality in health;	Identified key factors affecting data quality in health research and variability in assessment methods	Advocate for standardized data quality assessment methods in health research	4 - Bernardi et al. (2023). Data Quality in Health Research: Integrative Literature Review.
B - Propose the adoption of frameworks and tools, combined with traditional governance and information exchange techniques, to promote a data quality model for scientific research adopting the concept of digital health in their areas and research networks;	Developed a theoretical foundation for data quality in the context of RD	Implement and adapt the proposed data quality model across various health domains and settings	7- Bernardi et al. (2023). A proposal for a set of attributes relevant for Web portal data quality: The Brazilian Rare Disease Network case.
C - Propose a comprehensive set of data quality mechanisms;	Explored methods to support standardizing data sharing and reuse in RD scientific research	Recommend these approaches for broader application in health research to foster collaborative data use	5 - Bernardi et al. (2023). The Minimum Data Set for Rare Diseases: Systematic Review.
D - Develop a tool with guidelines for data collection, management, and evaluation in scientific research;	Highlighted challenges in TB data management and proposed FAIRification methods	Apply FAIR principles more broadly in TB data management and encourage the adoption of these practices in other health areas	8- Bernardi et al. (2023). From Raw Data to FAIR Data: The FAIRification Workflow for Brazilian Tuberculosis Research.
E - Propose approaches to support the sharing and reuse of data collected in scientific research;	Proposed a computational infrastructure for analyzing TB data and a model for health information management based on Lean thinking	Utilize the developed tools and models to enhance data management processes in other health research areas	9- Mozini et al. (2023). A Computational Infrastructure for Analyzing Tuberculosis Research Data in Brazil.;  10 - Teixeira et al. (2021) Proposal for a health information management model based on Lean thinking
F - Propose an implementation research model based on guiding principles and guidelines of digital health;	Applied the RE-AIM framework to evaluate the implementation of a health information system	Encourage the use of the RE-AIM framework as a standard for evaluating digital health interventions	11 - Crepaldi et al. (2023). An information system for monitoring tuberculosis cases: implementation research protocol using RE-AIM for a health region in Brazil.
To be continued			

Conclusion			
Objective/Hypothesis	Main Contributions	Recommendations	Related Chapter/ Publication
G - Promote and present the process of adherence, accreditation, and integration of a Brazilian research network in international research consortia engaged with implementation research principles.	Demonstrated effective integration of Brazilian research network with international consortia	Foster partnerships and collaborations with international consortia to share knowledge and resources in digital health research	6- Alves et al. (2021). Mapping, infrastructure, and data analysis for the Brazilian network of rare diseases: protocol for the RARASnet observational cohort study.

Source: The Author

In the first thematic arc, Chapters 4 and 5 by Bernardi et al. (2023) establish a theoretical foundation for understanding data quality in health research. The third chapter offers an integrative literature review on data quality in health research, setting a broad context for the subsequent discussions. Following this, the fourth chapter focuses specifically on RD, systematically reviewing the minimum data set necessary for research in this area.

This specific focus on RD is further expanded in Chapters 6 and 7, where Alves et al. (2021) and Bernardi et al. (2023) delve into the practical aspects of managing and analyzing data within the RARAS. Chapter 6 details the RARASnet observational cohort study, offering insights into the mapping, infrastructure, and data analysis techniques employed. In contrast, Chapter 7 proposes a set of attributes for web portal data quality, using the RARAS as a case study, thus applying the earlier theoretical concepts in a real-world context.

The second thematic arc shifts the focus to TB research in Brazil. Chapters 8 and 9, complement each other by covering the FAIRification workflow and computational infrastructure for TB research data. Chapter 8 outlines the process of transforming raw TB research data into FAIR data, highlighting the importance of data management in enhancing research efficiency and reproducibility. Meanwhile, Chapter 9 emphasizes the technical aspects of data analysis, providing a detailed look at the computational infrastructure designed to handle TB research data in Brazil. These chapters collectively showcase the application of data quality principles in another critical area of health research.

In Chapter 10, we introduce a novel perspective by proposing a health information management model based on Lean thinking, thus offering insights into optimizing health data management processes. Finally, Chapter 11 ties together the various aspects of health data management and quality discussed in previous chapters by detailing an information system for

monitoring TB cases using the RE-AIM framework in a practical setting. This chapter demonstrates the application of theoretical concepts in a real-world scenario and highlights the importance of a systematic approach to health data management and research.

The presented governance model emphasizes data quality optimization and is instrumental in laying a robust foundation for planning, executing, and evaluating health interventions. This approach is not only in line with the principles of ethical research but also optimizes the utilization of digital health innovations. Preconized by WHO, the model also harnesses the power of digital technologies and health innovation to accelerate the global attainment of health and well-being, specifically for TB and RD Brazilian populations.

The study operates under robust research governance structures, ensuring transparent data management and use, and offers opportunities for network collaboration with research partners to develop future works. Despite facing challenges like maintaining data quality and sustaining an integrated stakeholder engagement over time, the study's innovative approach presents promising prospects for advancing our understanding of the digital key elements able to reduce the lack of the translation of scientific evidence and the health real-world that deal with heterogeneous data.

Finally, the interdisciplinary nature of the research underscores its significant strength. By integrating insights from data science, public health, and health policy, the dissertation offers a holistic view of the challenges and opportunities in digital health. This collaborative approach is essential for addressing complex public health issues, where interdisciplinary perspectives can lead to more innovative and effective solutions. Integrating different methodologies and viewpoints in the dissertation highlights the importance of collaboration across disciplines in advancing digital health.

Within the realm of healthcare research, data governance stands as a cornerstone, playing a pivotal role in strengthening the foundation of robust data management. The thesis highlights the significance of data governance by emphasizing its role in ensuring data quality, security, and compliance with research standards, closely aligning with the core principles advocated throughout the research. Furthermore, the research underscores the importance of fostering collaboration and standardization within the healthcare research landscape. By facilitating cooperation among diverse stakeholders, including IT experts, research networks, and healthcare institutions, the study promotes a culture of shared practices, which is paramount for elevating the overall data quality.

## 13.2 Limitations

Throughout this doctoral thesis, a thorough examination of digital health in the context of RD and TB is conducted. Yet, it's important to acknowledge certain limitations referenced in the initial chapters. Chapters 4 and 5, which lay the theoretical foundation for understanding data quality in health research, have inherent limitations that stem primarily from the scope of the study and the specific focus of these chapters. Both chapters, while providing a solid theoretical base and contributing significantly to the field, are bound by the limitations of their respective scopes.

Although broad, the literature review in Chapter 4 may not capture the most current trends or unexplored areas in health data quality. Similarly, while detailed, the specific focus on RD in Chapter 5 restricts the applicability of its findings to other areas of health research. These limitations highlight the need for ongoing research and continuous literature updating to keep pace with the dynamic nature of health research and data quality standards.

In Chapters 6 and 7 of the doctoral thesis, which delve into the practical aspects of managing and analyzing data within the Brazilian network of RD, there are specific limitations related to the technological aspects and the rapidly evolving landscape of digital technologies. The limitations arise from the fast-paced advancements in data analysis, big data, and new digital health infrastructure and applications like blockchain and Large language models (LLM). The tools and methodologies employed, though state-of-the-art at the time, may quickly become outdated due to continuous technological innovations. This rapid evolution can potentially limit the long-term applicability and relevance of the findings and methodologies detailed in these chapters. This concern is also shared in Chapter 9.

In Chapter 8, which shifts the focus to TB research in Brazil, the limitations stem mainly from the challenges associated with data quality and management. The process of FAIRification relies heavily on the original quality of the raw data. Any inherent inaccuracies, biases, or incompleteness in the raw data could significantly impact the validity of the FAIRified data and, consequently, the research conclusions drawn from it.

Offering innovative approaches to health information management and TB case monitoring, also present limitations regarding the generalizability of their findings. While Chapters 10 and 11 introduce a health information management model inspired by Lean thinking provides valuable insights and could potentially enhance efficiency in health data management, its generalizability is limited. The context-specific nature of this model requires

careful consideration when applying it to different organizational cultures or healthcare infrastructures.

In this sense, demonstrates the practical application of theoretical concepts in a real-world setting, its findings and the developed system's applicability largely depend on the specific health region in Brazil where the study was conducted. The system's design and effectiveness might not be universally applicable to other regions or populations without significant adjustments, considering the variability in healthcare infrastructure, TB prevalence, and regional healthcare policies. This need for customization to specific contexts underscores a common challenge in health informatics research, where solutions often must be tailored to meet diverse healthcare needs and environments.

### **13.3 Future research**

Future research from this thesis should initially center on assessing the adaptability and efficacy of the governance model and data management strategies developed, particularly in the context of various diseases. Such research would involve a comprehensive investigation into the application of these models across both communicable and non-communicable diseases. This exploration is pivotal, as it could yield critical insights into the versatility and effectiveness of these models, thereby significantly contributing to the broader healthcare field. By offering scalable and adaptable solutions for diverse disease management, these studies have the potential to enhance overall healthcare outcomes substantially.

In parallel, exploring emerging data analysis technologies, such as artificial intelligence and machine learning, is essential. These advanced technologies hold considerable promise in refining the analysis of complex health data sets. Future studies should focus on integrating these technologies to enhance the analytical capabilities in health informatics. Moreover, applying predictive analytics to foresee disease trends and patient outcomes presents a proactive approach to public health management. This aspect of research could transform the reactive nature of current health systems into more anticipatory and responsive entities.

Another critical area for future exploration is integrating Brazilian health data systems with international counterparts. This research domain would require thoroughly examining the interoperability challenges and formulating strategies to align Brazilian health data systems with global standards. Research in this area is fundamental for facilitating international data sharing and collaboration, contributing to a more cohesive and unified approach to global health data management (WHO, 2021).

Furthermore, future research must delve into digital health technologies' policy and ethical implications. This includes a focused examination of these technologies' data privacy, security, and ethical use. Developing comprehensive guidelines or frameworks to navigate the complex ethical landscape of digital health research and practice is essential. Such frameworks would ensure the responsible and ethical use of technology in healthcare, addressing a critical need in the field (WILKISON *et al.*, 2016).

Longitudinal studies to evaluate the long-term impacts of the digital health strategies implemented in this research are also crucial. These studies would provide a more comprehensive understanding of the strategies' effectiveness over extended periods, contributing significantly to the evidence base supporting digital health interventions.

Additionally, investigating the user experience and acceptance of digital health tools among various stakeholders, including healthcare providers and patients, is vital. This research should focus on enhancing the design and functionality of digital health applications to improve usability and acceptance. Enhanced usability and acceptance are key to facilitating broader adoption and effectiveness of these tools.

Finally, the development of collaborative, multidisciplinary research projects that involve a diverse range of stakeholders stands as a crucial future direction. Such collaborative efforts promise to broaden the scope and depth of research, ensuring the integration of multiple perspectives and leading to more comprehensive and impactful outcomes in the field of digital health.

In conclusion, focusing on the scalability and sustainability of digital health interventions is imperative for future studies. This includes assessing the effectiveness of these interventions in larger populations or different geographic settings, as well as examining the financial, organizational, and technological factors that influence their long-term viability (GLASGOW *et al.*, 2019). These future research directions emanating from this thesis hold the potential to significantly advance the field of digital health, ensuring its relevance and applicability in improving global health outcomes.



## REFERENCES

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA. **Guia de inspeção em Boas Práticas Clínicas (BPC)**. Brasília: ANVISA, 2020a. Disponível em: <http://antigo.anvisa.gov.br/documents/10181/6023044/guia+36.pdf/290961fe-a808-4ad1-8ce3-fbaae6f43437>. Acesso em: 15 jan. 2024.

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA. **Nota técnica nº 23/2020/SEI/COPEC/GGMED/DIRE2/ANVISA**. Brasília: ANVISA, 2020b. Disponível em: <http://antigo.anvisa.gov.br/documents/219201/5923491/NOTA+T%C3%89CNICA+N%C2%BA+23-2020+-+GGMED.pdf/fc71b725-0457-43d3-aacc-4bceb3f4127f>. Acesso em: 15 jan. 2024.

ALKIRE, B. C. *et al.* The Economic consequences of mortality amenable to high-quality health care in low-and middle-income countries. **Health Affairs**, v. 37, n. 6, p. 988-996, 2018.

ARNIZANT, T. F. S. *et al.* My Latent tuberculosis treatment-mobile application to assist in adherence to latent tuberculosis treatment. **Procedia Computer Science**, v. 196, p. 640-646, 2022.

BAMMER, G. **Disciplining interdisciplinarity**: Integration and implementation sciences for researching complex real-world problems. Australia: ANU, 2013.

BASAJJA, M.; NAMBOBI, M.; WOLSTENCROFT, K. Possibility of enhancing digital health interoperability in Uganda through FAIR data. **Data Intelligence**, v. 4, n. 4, p. 899-916, 2022.

BERNARDI, F. *et al.* Blockchain based network for tuberculosis: a data sharing initiative in Brazil. **Studies in Health Technology and Informatics**, v. 262, p. 264-267, July 2019.

BERNARDI, F. A. *et al.* Mais do que palavras: uma análise das emoções brasileiras durante a COVID-19. **Revista Ibérica de Sistemas e Tecnologias de Informação**, n. E40, p. 526-541, 2021.

BERNARDI, F. A. *et al.* The Minimum dataset for rare diseases in Brazil: a systematic review protocol. **Procedia Computer Science**, v. 196, p. 439-444, 2022.

BRASIL. **Lei nº 9.507, de 12 de novembro de 1997**. Regula o direito de acesso a informações e disciplina o rito processual do habeas data. Disponível em: [http://www.planalto.gov.br/ccivil\\_03/Leis/L9507.htm](http://www.planalto.gov.br/ccivil_03/Leis/L9507.htm). Acesso em: 21 jan. 2022.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). **Diário Oficial União**, Brasília, DF, 15 ago. 2018, ed. 157, seção 1, p.59. Disponível em: [https://www.in.gov.br/materia/-/asset\\_publisher/Kujrw0TZC2Mb/content/id/36849373/do1-2018-08-15-lei-no-13-709-de-14-de-agosto-de-2018-36849337](https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/36849373/do1-2018-08-15-lei-no-13-709-de-14-de-agosto-de-2018-36849337). Acesso em: 27 jan. 2022.

BRASIL. **Lei geral de proteção aos pessoais:** guia de elaboração de termo de uso e política de privacidade para serviços públicos. 2020a. Disponível em: [https://www.gov.br/governodigital/pt-br/privacidade\\_e\\_seguranca/ppsi/guia\\_termo\\_uso\\_politica\\_privacidade.pdf](https://www.gov.br/governodigital/pt-br/privacidade_e_seguranca/ppsi/guia_termo_uso_politica_privacidade.pdf)  
Acesso em: 22 dez. 2023.

BRASIL. Ministério da Saúde. **Portaria nº 2.073, de 31 de agosto de 2011.** Brasília, DF: Ministério da Saúde, 2011. Disponível em: [https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073\\_3\\_1\\_08\\_2011.html](https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_3_1_08_2011.html).  
Acesso em: 16 dez. 2023.

BRASIL. Ministério da Saúde. **Cartão nacional de saúde.** Brasília, DF: Ministério da Saúde, 2015. Disponível em: <https://www.gov.br/saude/pt-br/aceso-a-informacao/acoes-e-programas/cartao-nacional-desauade>. Acesso em: 16 jan. 2022.

BRASIL. Ministério da Saúde. **Política nacional de informação e informática em saúde.** Brasília, DF: Ministério da Saúde, 2016. Disponível em: [https://bvsms.saude.gov.br/bvs/publicacoes/politica\\_nacional\\_informatica\\_saude\\_2016.pdf](https://bvsms.saude.gov.br/bvs/publicacoes/politica_nacional_informatica_saude_2016.pdf). Acesso em: 25 nov. 2021.

BRASIL. Ministério da Saúde. **DATASUS:** departamento de informática do sistema único de saúde - histórico. 2019a. Disponível em: <https://datasus.saude.gov.br/sobre-o-datasus>.  
Acesso em: 2 fev. 2022

BRASIL. Ministério da Saúde. **Sistema de informação de agravos de notificação - Sinan Net.** 2019b. Disponível em: <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/tubercsp.def>.  
Acesso em: 6 fev. 2022.

BRASIL. Ministério da Saúde. **Plano de ação, monitoramento e avaliação da estratégia de saúde digital para o Brasil 2019-2023.** Brasília, DF: Ministério da Saúde, 2020b. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-digital/a-estrategiabrasileira/>.  
Acesso em: 27 jan. 2022.

BRASIL. Ministério da Saúde. **Portaria Nº 1.434, de 28 de maio de 2020.** Institui o Programa Conecte SUS e altera a Portaria de Consolidação nº 1/GM/MS, de 28 de setembro de 2017, para instituir a Rede Nacional de Dados em Saúde e dispor sobre a adoção de padrões de interoperabilidade em saúde. Brasília, DF: Ministério da Saúde, 2020c. Disponível em: <https://www.in.gov.br/en/web/dou/-/portaria-n1.434-de-28-de-maio-de-2020-259143327>. Acesso em: 15 dez. 2021.

BRASIL. Ministério da Saúde. **Metodologia de desenvolvimento de software MDS.** Disponível em: <http://datasus.saude.gov.br/metodologias-mds/>. Acesso em 4. fev. 2022.

BRASIL. Ministério da Saúde. Organização Pan-Americana da Saúde. Fundação Oswaldo Cruz. **A Experiência brasileira em sistemas de informação em saúde.** Brasília, DF: Ed.Ministério da Saúde, 2009. Disponível em: [http://bvsms.saude.gov.br/bvs/publicacoes/experiencia\\_brasileira\\_sistemas\\_saude\\_volume1.pdf](http://bvsms.saude.gov.br/bvs/publicacoes/experiencia_brasileira_sistemas_saude_volume1.pdf). Acesso em: 6 fev. 2022.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Especializada e Temática. Coordenação Geral de Média e Alta Complexidade. **Diretrizes doenças raras**: portaria GM/MS nº 199 de 30/01/2014. Brasília, DF: Ministério da Saúde, 2014. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/protocolos-clinicos-e-diretrizes-terapeuticas-pcdt/arquivos/2014/diretrizes-doencas-raras.pdf/view>. Acesso em: 16 fev. 2022.

BRASIL. Ministério da Saúde. Secretaria Executiva - DATASUS. **Plano diretor de tecnologia da informação 2019-20121**. Brasília, DF: Ministério da Saúde, 2020d. Disponível em: <https://datasus.saude.gov.br/wpcontent/uploads/2020/05/22052020v5.pdf>. Acesso em: 14 nov. 2021.

BRASIL. Ministério da Saúde. Secretária Executiva. Departamento de Informática do SUS. **Estratégia de saúde digital para o Brasil 2020-2028**. Brasília, DF: Ministério da Saúde; Secretária Executiva; Departamento de Informática do SUS, 2020e. Disponível em: [https://bvsmms.saude.gov.br/bvs/publicacoes/estrategia\\_saude\\_digital\\_Brasil.pdf](https://bvsmms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf). Acesso em: 15 jan. 2022

BRASIL. Ministério da Ciência, Tecnologia e Inovações. Gestão e governança de dados. 2020f. Disponível em: <https://www.gov.br/mcti/pt-br/coggd/gestao-e-governanca-de-dados>. Acesso em: 22 jan. 2022.

CANADA HEALTH INFOWAY. **Modelos internacionais de governança em saúde digital**: pesquisa. 2016. Disponível em: <https://pesquisa.bvsalud.org/portal/resource/pt/biblio-1348473>. Acesso em: 16 nov. 2023.

CARVALHO, I. *et al.* The Collaborative data gathering and audit process in the Covid-19 Brazil portal: health information in the pandemic. **Revista de Saúde Digital e Tecnologias Educacionais**, v. 7, n. 1, 2022a. Disponível em: [http://periodicos .ufc.br/resdite/index](http://periodicos.ufc.br/resdite/index). Acesso em: 15 jan. 2024.

CARVALHO, I. *et al.* COVID-19 BR: a web portal for COVID-19 information in Brazil. **Procedia Computer Science**, v. 196, p. 525-532, 2022b.

CARVALHO, I. *et al.* Knowledge discovery in databases: comorbidities in tuberculosis cases. *In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE*, 22., 2022, London. **Proceedings [...]**. Berlim: Springer Nature, 2022c. p. 3-13.

CASSÃO, V. *et al.* Unsupervised analysis of COVID-19 pandemic evolution in brazilian states. **Procedia Computer Science**, v. 196, p. 655-662, 2022.

CASSÃO, V. *et al.* A Web portal for real-time data quality analysis on the brazilian tuberculosis research network: a case study. *In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE*, 23., 2023, Prague. **Proceedings [...]**. Berlim: Springer Nature, 2023. p. 300-312.

COSTA, L. M. A. *et al.* Operational modeling for testing diagnostic tools impact on tuberculosis diagnostic cascade: a model design. **Procedia Computer Science**, v. 181, p. 650-657, 2021.

DAVIS, S. L. M. *et al.* A Democracy deficit in digital health?. **Health and Human Rights Journal**, Jan. 2020. Disponível em: <https://www.hhrjournal.org/2020/01/a-democracy-deficit-in-digital-health>. Acesso em: 22 jan. 2024.

DILLON, D. G. *et al.* Open-source electronic data capture system offered increased accuracy and cost-effectiveness compared with paper methods in Africa. **Journal of Clinical Epidemiology**, v. 67, n. 12, p. 1358-1363, 2014.

FAIRsharing - a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies. Oxford: University of Oxford, 2011. Disponível em: [https://lib.uva.nl/discovery/fulldisplay?vid=31UKB\\_UAM1\\_INST:UVA&search\\_scope=DN\\_and\\_CI\\_and\\_PURE&docid=alma9940750682405131](https://lib.uva.nl/discovery/fulldisplay?vid=31UKB_UAM1_INST:UVA&search_scope=DN_and_CI_and_PURE&docid=alma9940750682405131). Acesso em: 28 dez. 2021.

FÉLIX, T. M. *et al.* Epidemiology of rare diseases in Brazil: protocol of the brazilian rare diseases network (RARAS-BRDN). **Orphanet Journal of Rare Diseases**, v. 17, n. 1, p. 84, 2022. DOI: 10.1186/s13023-022-02254-4.

FOOD AND DRUG ADMINISTRATION. **Digital health innovation action plan**. 2019. Disponível em: <https://www.fda.gov/media/106331/download>. Acesso em: 22 jan. 2023.

GALVÃO, M. C. B.; PLUYE, P.; RICARTE, I. L. M. Métodos de pesquisa mistos e revisões de literatura mistas: conceitos, construção e critérios de avaliação. **InCID: Revista de Ciência da Informação e Documentação**, v. 8, n. 2, p. 4-24, 2017.

GLASGOW, R. E.; VOGT, T. M.; BOLES, S. M. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. **American Journal of Public Health**, v. 89, n. 9, p. 1322-1327, 1999.

GLASGOW, R. E. *et al.* RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. **Frontiers in Public Health**, v. 7, p. 64, 2019. DOI: 10.3389/fpubh.2019.00064.

GO FAIR. **How to GO FAIR**. 2022. Disponível em: <https://www.go-fair.org/how-to-go-fair/>. Acesso em: 22 jan. 2022.

HARRISON, K.; RAHIMI, N.; DANOVARO-HOLLIDAY, M. C. Factors limiting data quality in the expanded programme on immunization in low and middle-income countries: a scoping review. **Vaccine**, v. 38, n. 30, p. 4652-4663, 2020.

HARZHEIM, E. *et al.* **Guia de avaliação, implantação e monitoramento de programas e serviços em telemedicina e telessaúde**. Porto Alegre: Universidade Federal do Rio Grande do Sul; Hospital Alemão Oswaldo Cruz, 2017.

HERSH, W. Health care information technology: progress and barriers. **Jama**, v. 292, n. 18, p. 2273-2274, 2004.

KERNEBECK, S. *et al.* Adherence to digital health interventions: definitions, methods, and open questions. **Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz**,

v. 64, n. 10, p. 1278-1284, 2021.

KHARE, R. *et al.* A Longitudinal analysis of data quality in a large pediatric data research network. **Journal of the American Medical Informatics Association**, v. 24, n. 6, p. 1072-1079, 2017.

KHOR, L. *et al.* **Content creator guidelines**: how to manage digital surveys for aquaculture using KoBoToolbox and ODK collect mobile app. 2020. Disponível em: <https://digitalarchive.worldfishcenter.org/handle/20.500.12348/4892>. Acesso em: 23 fev. 2022.

KLIPIN, M. *et al.* The Process of installing REDCap, a web based database supporting biomedical research. **Applied Clinical Informatics**, v. 5, n. 4, p. 916-929, 2014.

KNIBERG, H. **Scrum and XP from the trenches**. 2<sup>nd</sup>ed. London: InfoQ Newsletter, 2015.

KOSTKOVA, P. Grand challenges in digital health. **Frontiers in Public Health**, v. 3, May 2015. DOI: 10.3389/fpubh.2015.00134.

LABRIQUE, A. *et al.* WHO digital health guidelines: a milestone for global health. **NPJ Digital Medicine**, v. 3, n. 1, p. 1-3, 2020.

LIMA, V. C. *et al.* A Permissioned blockchain network for security and sharing of de-identified tuberculosis research data in Brazil. **Methods of Information in Medicine**, v. 59, n. 6, p. 205-218, 2020.

LIMA, V. C. *et al.* Security framework for tuberculosis health data interoperability through the semantic WEB. **International Journal of Web Portals**, v. 13, n. 2, p. 36–57, 2021a.

LIMA, V. C. *et al.* **REDBOX**: a comprehensive semantic framework for data collection and management in tuberculosis research. 2021b. DOI: <https://doi.org/10.21203/rs.3.rs-1070973/v1>.

LIMA, V. C. *et al.* The Relevance of digital health strategies to support tuberculosis services in the COVID-19 pandemic context. **Journal of Multiprofessional Health Research**, v. 2, n. 1, p. e02.72-e02.74, 2021c.

LIMA, V. C. *et al.* Security framework for tuberculosis health data interoperability through the semantic web. **International Journal of Web Portals (IJWP)**, v. 13, n. 2, p. 36-57, 2021d.

LIMA, V. C. *et al.* A Mechanism for verifying the integrity and immutability of tuberculosis data using IOTA distributed ledger technology. *In*: HAYN, D.; SCHREIER, G.; BAUMGARTNER, M. (Ed.). **Navigating healthcare through challenging times**. Amsterdam: IOS, 2021e. p. 130-135. (SStudies in Health Technology and Information, 279).

LIMA, V. C. *et al.* A Computational infrastructure for semantic data integration towards a patient-centered database for tuberculosis care. **Procedia Computer Science**, v. 196, p. 434-438, 2022.

LIU, J. *et al.* Epidemiological, clinical characteristics and outcome of medical staff infected with COVID-19 in Wuhan, China: a retrospective case series analysis. **MedRxiv**, 2020. DOI: <https://doi.org/10.1101/2020.03.09.20033118>.

MACIEL, D. A.; FERREIRA, D. P.; MARIN, H. F. Padrões de terminologias nacionais para procedimentos e intervenções na saúde. **Revista de Administração em Saúde**, v. 18, n. 71, abr./jun. 2018. DOI: <http://dx.doi.org/10.23973/ras.71.111>.

MARTINS, P. E. A. *et al.* Supervised machine learning techniques applied to medical records toward the diagnosis of rare autoimmune diseases. *In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE*, 23., 2023, Prague. **Proceedings [...]**. Berlim: Springer Nature, 2023. p. 170-184.

MESKÓ, B. *et al.* Digital health is a cultural transformation of traditional healthcare. **Mhealth**, v. 3, Sept. 2017. DOI: [10.21037/mhealth.2017.08.07](https://doi.org/10.21037/mhealth.2017.08.07).

METKE-JIMENEZ, A.; HANSEN, D. FHIRCap: Transforming REDCap forms into FHIR resources. **AMIA Joint Summits on Translational Science Proceedings**, p.54-63, May 2019. PMC6568121.

MÖLLER, M. *et al.* Representing the international classification of diseases version 10 in OWL. *In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND ONTOLOGY DEVELOPMENT - KEOD*, 2010, Valência. **Proceedings [...]**. Setúbal: SciTePress, 2010. p. 50-59.

MONTINI, N. R. **Necessidade imediata de consolidação da ANPD para efetividade da LGPD e prevenção de excessivas demandas judiciais**. 2020. Monografia (Trabalho de Conclusão de Curso) – Faculdade de Ciências Jurídicas e Sociais, Centro universitário de Brasília, 2020. Disponível em: <https://repositorio.uniceub.br/jspui/handle/prefix/14664>. Acesso em: 15 jan. 2022.

MOURA JÚNIOR, L. A. A Estratégia de saúde digital para o Brasil 2020-2028. **Journal of Health Informatics**, v. 13, n. 1, p.1-2, 2021.

NATIONAL INSTITUTE OF HEALTH. **NIH data sharing policy and implementation guidance**. 2021. Disponível em: [https://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_guidance.htm#app](https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#app). Acesso em: 22 jan. 2022.

NEIVA, M. B. *et al.* Brazil: the emerging epicenter of COVID-19 pandemic. **Revista da Sociedade Brasileira de Medicina Tropical**, v.53, 2020. DOI: <https://doi.org/10.1590/0037-8682-0550-2020>.

NEIVA, M. B. *et al.* Informed consent form automated validation, the brazilian rare disease network case proposal. **Procedia Computer Science**, v. 219, p. 1538-1545, 2023.

OH, H. *et al.* What is eHealth (3): a systematic review of published definitions. **Journal of Medical Internet Research**, v. 7, n. 1, Jan./Mar. 2005. DOI: [10.2196/jmir.7.1e1](https://doi.org/10.2196/jmir.7.1e1).

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. **Boas práticas clínicas: documento das Américas**. 2005. Disponível em:

[https://bvsm.s.saude.gov.br/bvs/publicacoes/boas\\_praticas\\_clinicas\\_opas.pdf](https://bvsm.s.saude.gov.br/bvs/publicacoes/boas_praticas_clinicas_opas.pdf).

Acesso em: 22 jan. 2022.

PASALIC, D. *et al.* Implementing an electronic data capture system to improve clinical workflow in a large academic radiation oncology practice. **JCO Clinical Cancer Informatics**, v. 2, p. 1-12, 2018.

PELLISON, F. C. *et al.* Data integration in the Brazilian public health system for tuberculosis: use of the semantic web to establish interoperability. **JMIR Medical Informatics**, v. 8, n. 7, July 2020. DOI: 10.2196/17176.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: an update. **Information and Software Technology**, v. 64, p. 1-18, 2015.

POPPE, O.; SÆBØ, J. I.; BRAA, J. WHO digital health packages for disseminating data standards and data use practices. **International Journal of Medical Informatics**, v. 149, May 2021. DOI: 10.1016/j.ijmedinf.2021.104422.

PULLONEN, P. *et al.* Privacy-enhanced BPMN: enabling data privacy analysis in business processes models. **Software and Systems Modeling**, v. 18, n. 6, p. 3235-3264, 2019.

RAMOS, A. B.; VILELA JUNIOR, D. C. A Influência do papel do scrum master no desenvolvimento de projetos Scrum. **Revista de Gestão e Projetos**, v. 8, n. 3, p. 80-99, 2017.

REIS, M. *et al.* Development of a mobile application with health guidelines for TB patients care. **Stud Health Technol Inform**, v. 305, p. 373-376, 2023. DOI: 10.3233/SHTI230509.

RICHTER, T. *et al.* Rare disease terminology and definitions—a systematic global review: report of the ISPOR rare disease special interest group. **Value in Health**, v. 18, n. 6, p. 906-914, 2015.

ROWLANDS, D. What is digital health? And why does it matter. Australia: Digital Health Workforce Academy; Australian Institute of Digital Health, 2019.

SANSONE, S.; MCQUILTON, P.; ROCCA-SERRA, P. Fairsharing as a community approach to standards, repositories and policies. **Nature Biotechnology**, v. 37, p. 358–367, 2019. Disponível em: <https://www.nature.com/articles/s41587-019-0080-8>. Acesso em: 28 dez. 2021.

SANTOS, A. C. *et al.* **Sistema de informações hospitalares do Sistema Único de Saúde: documentação do sistema para auxiliar o uso das suas informações**. 2009. Dissertação (Mestrado) – Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, 2009.

SANTORO, E. Information technology and digital health to support health in the time of

CoViD-19. **Recenti Progressi in Medicina**, v. 111, n. 7, p. 393-397, 2020.

SANTOS, W. S. *et al.* Reflexões acerca do uso da telemedicina no Brasil: oportunidade ou ameaça?. **Revista de Gestão em Sistemas de Saúde**, v. 9, n. 3, p. 433-453, 2020.

SOCIEDADE BRASILEIRA DE INFORMATICA EM SAÚDE. História da SBIS. 2020. Disponível em: <http://www.sbis.org.br/historia-da-sbis>. Acesso em: 15 fev. 2022.

SILVA, E. A. Evolução histórica do método científico: desafios e paradigmas para o século XXI. **Revista Economia & Pesquisa**, v. 3, n. 3, p. 109-118, mar. 2001.

SILVA, A. B.; MORAES, I. H. S. O Caso da rede universitária de telemedicina: análise da entrada da telessaúde na agenda política brasileira. **Physis: revista de saúde coletiva**, v. 22, n. 3, p. 1211-1235, 2012.

SINGHAL, A.; SRIVASTAVA, J. Generating semantic annotations for research datasets. *In: INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS*, 4., 2014, Thessaloniki Greece. **Proceedings [...]**. New York: Association for Computing Machinery, 2014.

DOI: <https://doi.org/10.1145/2611040.2611056>.

SOARES, E. V. B. **Atenção básica e informação: análise do sistema de informação em saúde para atenção básica (SISAB) e estratégia e-SUS AB e suas repercussões para uma gestão da saúde com transparência**. Brasília: Universidade de Brasília; Faculdade de Economia, Administração e Contabilidade, Departamento de Administração, 2016. Artigo apresentado ao Departamento de Administração como requisito parcial à obtenção do título de Especialista em Gestão Pública na Saúde.

SOARES, G. T. *et al.* Scaling laws and spatial effects of Brazilian health regions: a research protocol. **Procedia Computer Science**, v. 219, p. 1325-1332, 2023.

SWERTZ, M. *et al.* Towards an interoperable ecosystem of research cohort and real-world data catalogues enabling multi-center studies. **Yearbook of Medical Informatics**, v. 31, n. 1, p. 262-272, 2022.

TILAHUN, B.; FRITZ, F. Comprehensive evaluation of electronic medical record system use and user satisfaction at five low-resource setting hospitals in Ethiopia. **JMIR Medical Informatics**, v. 3, n. 2, 2015. DOI: 10.2196/medinform.4106.

US DEPARTMENT OF HEALTH AND HUMAN SERVICES. Office for Human Research Protections. Office of the Assistant Secretary for Health. **International compilation of human research standards**. 2021. Disponível em:

<https://www.hhs.gov/ohrp/international/compilation-human-research-standards/index.htm>.

Acesso em: 15 fev. 2022.

VÄRRI, A. *et al.* **Integrated citizen centered digital health and social care: citizens as data producers and service co-creators**. Amsterdam: IOS, 2020. (Studies in Health Technology and Informatics, 275).

WANG, L. X.; BLOCH, C. Digital Interventions in the Health Sector—Country Cases and Policy Discussions. Washington: International Monetary Found, 2023. (Note/2023/004).



WATTS, G. Unicorns and cowboys in digital health: the importance of public perception. **Lancet Digital Health**, v. 1, n. 7, 2019. DOI: 10.1016/S2589-7500(19)30164-5.

WEISKOPF, N. G.; WENG, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. **Journal of the American Medical Informatics Association**, v. 20, n. 1, p. 144-151, 2013.

WILKINSON, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 1, p. 1-9, 2016.

WORLD HEALTH ORGANIZATION. **Implementation research for the control of infectious diseases of poverty**: strengthening the evidence base for the access and delivery of new and improved tools, strategies and interventions. Geneva: WHO; TDR, 2011.

WORLD HEALTH ORGANIZATION. **National eHealth strategy toolkit**. Geneva: WHO; International Telecommunication Union, 2012.

WORLD HEALTH ORGANIZATION. **Engage-TB**: integrating community-based tuberculosis activities into the work of nongovernmental and other civil society organizations: implementation manual. Geneva: WHO, 2013.

WORLD HEALTH ORGANIZATION. **A Global action framework for TB research in support of the third pillar of WHO's end TB strategy**. Geneva: WHO, 2015.

WORLD HEALTH ORGANIZATION. **Framework on integrated, people-centred health services**. Geneva: WHO, 2016. (Sixty-Ninth World Health Assembly: provisional agenda item 16.1 – A69/39).

WORLD HEALTH ORGANIZATION. **Classification of digital health interventions v1.0**: a shared language to describe the uses of digital technology for health. Geneva: WHO, 2018a.

WORLD HEALTH ORGANIZATION. **Technical report on critical concentrations for drug susceptibility testing of medicines used in the treatment of drug-resistant tuberculosis**. Geneva: WHO, 2018b.

WORLD HEALTH ORGANIZATION. **WHO guideline**: recommendations on digital interventions for health system strengthening. Geneva: WHO, 2019a.

WORLD HEALTH ORGANIZATION. **TDR implementation research toolkit**. Geneva: WHO, 2019b.

WORLD HEALTH ORGANIZATION. **Global strategy on digital health 2020-2025**. Geneva: WHO, 2021.

WRIGHT, A. REDCap: a tool for the electronic capture of research data. **Journal of Electronic Resources in Medical Libraries**, v. 13, n. 4, p. 197-201, 2016.

YAMADA, D. B. *et al.* National network for rare diseases in Brazil: the computational infrastructure and preliminary results. *In: GROEN, D. et al. (Ed.). Computational Sciences -International Conference on Computational Science 2022.* Cham: Springer International, 2022. p. 43-49. (Lecture Notes in Computer Science, 13352).

ZHANG, A. Quality improvement through Poka–Yoke: from engineering design to information system design. **International Journal of Six Sigma and Competitive Advantage**, v. 8, n. 2, p. 147-159, 2014.

# APPENDIX A – BRAZILIAN RARE DISEASE PORTAL COMPUTER PROGRAM REGISTRATION CERTIFICATES



REPÚBLICA FEDERATIVA DO BRASIL  
 MINISTÉRIO DA ECONOMIA  
**INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL**  
 DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

## Certificado de Registro de Programa de Computador

Processo Nº: **BR512022003289-9**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 28/02/2022, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

**Título:** Portal Brasileiro de Doenças Raras

**Data de publicação:** 28/02/2022

**Data de criação:** 17/03/2020

**Titular(es):** DOMINGOS ALVES; FILIPE ANDRADE BERNARDI; TÊMIS MARIA FELIX

**Autor(es):** DIEGO BETTIOL YAMADA; BIBIANA MELLO DE OLIVEIRA; VINÍCIUS COSTA LIMA; MARIANE BARROS NEIVA; MÁRCIO ELOI COLOMBO FILHO; ANDRÉ LUIZ TEIXEIRA VINCI; GIOVANE THOMAZINI SOARES; YASMIN DE ARAÚJO RIBEIRO; ISABELLE CARVALHO

**Linguagem:** HTML; JAVA SCRIPT; PYTHON; PHP; MYSQL; CSS; R

**Campo de aplicação:** BL-02; IF-04; IF-07; IF-09; IF-10; SD-08; SD-09

**Tipo de programa:** AP-01; AP-02; AP-03; AP-04; AT-06; DS-07; FA-01; FA-03; FA-04; GI-01; GI-02; GI-03; GI-04; GI-05; GI-06; GI-07; GI-08; PD-04; PD-05; TC-01; TC-02

**Algoritmo hash:** SHA-512

**Resumo digital hash:**

cade575796c9bdfca10bc136f5bfe6750e268977460f3494e649be88e507e74096e93e569926d6498b3a9ad992431ec178d93d23750c6b8e56c9ecebfb3d62b

**Expedido em:** 06/12/2022

**Aprovado por:**

Carlos Alexandre Fernandes Silva  
 Chefe da DIPTO

## APPENDIX B – SISTB COMPUTER PROGRAM REGISTRATION CERTIFICATES



REPÚBLICA FEDERATIVA DO BRASIL  
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS  
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL  
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS

### Certificado de Registro de Programa de Computador

Processo Nº: **BR512023002666-2**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 01/01/2013, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

**Título:** SISTB - Sistema da Tuberculose

**Data de publicação:** 01/01/2013

**Data de criação:** 01/01/2012

**Titular(es):** DOMINGOS ALVES; VINÍCIUS COSTA LIMA; FILIPE ANDRADE BERNARDI; NATHALIA YUKIE CREPALDI; FACULDADE DE MEDICINA DE RIBEIRÃO PRETO - UNIVERSIDADE DE SÃO PAULO

**Autor(es):** VINÍCIUS COSTA LIMA; NATHALIA YUKIE CREPALDI

**Linguagem:** HTML; JAVA; PHP; CSS

**Campo de aplicação:** CO-02; IF-01; IF-02; IN-02; SD-01; SD-03

**Tipo de programa:** AP-01; GI-01; TC-01

**Algoritmo hash:** SHA-512

**Resumo digital hash:**

20af75d1285c7b80e82d17b7503607edcef0613e4f445d2032d56ccae2b912ac7b428661dad8082fcaea230044ce5092b698830f5533c88fd2a0a4fc9e01fe98

**Expedido em:** 19/09/2023

**Aprovado por:**  
Carlos Alexandre Fernandes Silva  
Chefe da DIPTO