UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

FAIR Data in Medical Research

Incorporating the FAIR Principles in the Research Data Life Cycle

Kersloot, M.G.

Publication date
2022

Link to publication

Citation for published version (APA):
Kersloot, M. G. (2022). *FAIR Data in Medical Research: Incorporating the FAIR Principles in the Research Data Life Cycle*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:23 Jan 2025

Appendices

References

[1] Molloy JC. The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biology. 2011 Dec;9(12):e1001195.

[2] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996 Jan;312(7023):71–72.

[3] Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. BMJ. 1995 Apr;310(6987):1122–1126.

[4] Dammann O, Smart B. 5. In: Making Population Health Knowledge. Cham: Springer International Publishing; 2019. p. 63–77.

[5] Whyte A, Tedds J. Making the Case for Research Data Management. DCC Briefing Papers. 2011 09;.

[6] Corti L, Van den Eynden V, Bishop L, Woollard M. 2. In: The Research Data Lifecycle. Sage; 2019. .

[7] Jisc. Research data management toolkit; 2021.

[8] Pinnock G. The Research Data Management (RDM) lifecycle at the University of Cape Town (UCT); 2019.

[9] van Valkenhoef G, Tervonen T, de Brock B, Hillege H. Deficiencies in the transfer and availability of clinical trials evidence: a review of existing systems and standards. BMC Medical Informatics and Decision Making. 2012 Sep;12(1).

[10] Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE. 2015 Aug;10(8):e0134826.

[11] Committee on Strategies for Responsible Sharing of Clinical Trial Data IoM Board on Health Sciences Policy. Sharing clinical trial data: maximizing benefits, minimizing risk. National Academies Press Washington, DC; 2015.

[12] Shaw DL, Ross JS. US Federal Government Efforts to Improve Clinical Trial Transparency with Expanded Trial Registries and Open Data Sharing. AMA Journal of Ethics. 2015 Dec;17(12):1152–1159.

[13] Borgman CL. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology. 2012 Apr;63(6):1059–1078.

[14] Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Mithu Lucraft, et al. Whitepaper: Practical challenges for researchers in data sharing. 2018;.

[15] Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, et al. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science. 2015;.

[16] Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use. 2017 Mar;37(1):49–56.

[17] Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today. 2019 Apr;24(4):933–938.

[18] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016 Mar;3(1).

[19] Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, et al. A Generic Workflow for the Data FAIRification Process. Data Intelligence. 2020 Jan;2(1-2):56–65.

[20] GO FAIR. FAIRification Process; 2019.

[21] Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? European Journal of Human Genetics. 2018 May;26(7):931–936.

[22] Gruber TR. A translation approach to portable ontology specifications. Knowl Acquis. 1993;5.

[23] Creative Commons. CC0 licence; 2020.

[24] Reisen M, Oladipo F, Stokmans M, Mpezamihgo M, Folorunso S, Schultes E, et al. Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced Genetics. 2021 Jun;2(2).

[25] Bloemers M, Montesanti A. The FAIR Funding Model: Providing a Framework for Research

Funders to Drive the Transition toward FAIR Data Management and Stewardship Practices. Data Intelligence. 2020 01;2(1-2):171–180.

[26] National Institutes of Health - OSC: The Common Fund. NIH Data Commons Pilot Phase Consortium; 2018.

[27] Directorate-General for Research and Innovation (European Commission). Turning FAIR into reality. European Commission; 2018.

[28] Jones S, Pergl R, Hooft R, Miksa T, Samors R, Ungvari J, et al. Data Management Planning: How Requirements and Solutions are Beginning to Converge. 2020 Jan;2(1-2):208–219.

[29] for Research ECDG, Innovation, Services PE. Cost-benefit analysis for FAIR research data: cost of not having FAIR research data. Publications Office; 2018.

[30] Choudhury A, van Soest J, Nayak S, Dekker A. Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Bhattacharjee A, Borgohain SK, Soni B, Verma G, Gao XZ, editors. Machine Learning, Image Processing, Network Security and Data Sciences. Singapore: Springer Singapore; 2020. p. 85–95.

[31] SURF. FAIR data advanced use cases: from principles to practice in The Netherlands; 2018.

[32] Kaliyaperumal R, Wilkinson MD, Alarcón Moreno P, Benis N, Cornet R, dos Santos Vieira B, et al. Semantic modelling of Common Data Elements for Rare Disease registries, and a prototype workflow for their deployment over registry data. medRxiv. 2021;.

[33] Moreira JLR, Bonino L, Ferreira Pires L, Van Sinderen M, Henning P. Towards Findable, Accessible, Interoperable and Reusable (FAIR) Data Repositories: Improving a Data Repository to Behave as a FAIR Data Point | Repositórios para dados localizáveis, acessíveis, interoperáveis e reutilizáveis (FAIR): adaptando um repositór. Liinc em Revista. 2019;15(2):244–258.

[34] Sinaci AA, Núñez-Benjumea FJ, Gencturk M, Jauer ML, Deserno T, Chronaki C, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods of Information in Medicine. 2020 Jun;59(S 01):e21–e32.

[35] Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, Long Q, Joosten SA, van der Wijk HJ, et al. Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. 2021 Aug;.

[36] Kochev N, Jeliazkova N, Paskaleva V, Tancheva G, Iliev L, Ritchie P, et al. Your Spreadsheets Can Be FAIR: A Tool and FAIRification Workflow for the eNanoMapper Database. 2020 Sep;10(10):1908.

[37] Mons B. Data Stewardship for Open Science. Chapman and Hall/CRC; 2018.

[38] Scholtens S, Jetten M, Böhmer J, Staiger C, Slouwerhof I, van der Geest M, et al. Final report: Towards FAIR data steward as profession for the lifesciences. Report of a ZonMw funded collaborative approach built on existing expertise; 2019.

[39] Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence. 2020 Jan;2(1-2):10–29.

[40] Ford E, Nicholson A, Koeling R, Tate A, Carroll J, Axelrod L. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013;13.

[41] Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Informatics Assoc. 2011;18.

[42] van Reisen M, Stokmans M, Basajja M, Ong'ayo AO, Kirkpatrick C, Mons B. Towards the Tipping Point for FAIR Implementation. Data Intelligence. 2020 Jan;2(1-2):264–275.

[43] Thompson M, Burger K, Kaliyaperumal R, Roos M, da Silva Santos LOB. Making FAIR Easy with FAIR Tools: From Creolization to Convergence. Data Intelligence. 2020 Jan;2(1-2):87–95.

[44] European Commission. Guidelines on Data Management in Horizon 2020; 2013.

[45] The Netherlands Organisation for Scientific Research (NWO). Data management protocol;.

[46] LUMC. Research ICT; 2016.

[47] Radboud University Research Data Management. FAIR principles; 2019.

[48] Vesteghem C, Brøndum RF, Sønderkær M, Sommer M, Schmitz A, Bødker JS, et al. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Briefings in Bioinformatics. 2019 Jun;21(3):936–945.

[49] Joukes E, Cornet R, de Bruijne MC, de Keizer NF, Abu-Hanna A. Development and validation of a model for the adoption of structured and standardised data recording among healthcare professionals. BMC Medical Informatics and Decision Making. 2018 Jun;18(1).

[50] Castor EDC. Castor Electronic Data Capture; 2021.

[51] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020.

[52] Sarstedt M, Ringle CM, Hair JF. Partial Least Squares Structural Equation Modeling. In: Handbook of Market Research. Springer International Publishing; 2017. p. 1–40.

[53] Cassel C, Hackl P, Westlund AH. Robustness of partial least-squares method for estimating latent variable quality structures. Journal of Applied Statistics. 1999 May;26(4):435–446.

[54] Sanchez G, Trinchera L, Russolillo G. plspm: Tools for Partial Least Squares Path Modeling (PLS-PM); 2017. R package version 0.4.9.

[55] Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. European Business Review. 2019 Jan;31(1):2–24.

[56] Hair Jr JF, Hult GTM, Ringle C, Sarstedt M. A primer on partial least squares structural equation modeling (PLS-SEM). Sage publications; 2016.

[57] Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science. 2014 Aug;43(1):115–135.

[58] Trifan A, Oliveira JL. Towards a More Reproducible Biomedical Research Environment: Endorsement and Adoption of the FAIR Principles. In: Biomedical Engineering Systems and Technologies. Springer International Publishing; 2020. p. 453–470.

[59] Zuiderwijk A, Shinde R, Jeng W. What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. PLOS ONE. 2020 Sep;15(9):e0239283.

[60] Wilkinson MD, Sansone SA, Schultes E, Doorn P, da Silva Santos LOB, Dumontier M. A design framework and exemplar metrics for FAIRness. Scientific Data. 2018 Jun;5(1).

[61] Archiving D, Services N. Self-Assessment Tool to Improve the FAIRness of Your Dataset; 2019.

[62] Jannik S, Dennis K, Jens G, Christian-Alexander B, Marco R, van Enckevort David, et al. OSSE Goes FAIR - Implementation of the FAIR Data Principles for an Open-Source Registry for Rare Diseases. Studies in Health Technology and Informatics. 2018;253(German Medical Data Sciences: A Learning Healthcare System):209–213.

[63] Groenen KHJ, Jacobsen A, Kersloot MG, dos Santos Vieira B, van Enckevort E, Kaliyaperumal R, et al. The de novo FAIRification process of a registry for vascular anomalies. 2021 Sep;16(1).

[64] Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intelligence. 2020 Jan;2(1-2):96–107.

[65] Schultes E, Magagna B, Hettne KM, Pergl R, Suchánek M, Kuhn T. Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Lecture Notes in Computer Science. Springer International Publishing; 2020. p. 138–147.

[66] van Vlijmen H, Mons A, Waalkens A, Franke W, Baak A, Ruiter G, et al. The Need of Industry to Go FAIR. Data Intelligence. 2020 Jan;2(1-2):276–284.

[67] Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C. Electronic health records: new opportunities for clinical research. J Intern Med. 2013;274.

[68] Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014;52.

[69] Price SJ, Stapley SA, Shephard E, Barraclough K, Hamilton WT. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case–control study. BMJ Open. 2016 May;6(5):e011664.

[70] SNOMED International. SNOMED CT; 2017.

[71] Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP. Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47.

[72] Krasowski M, Schriever A, Mathur G, Blau J, Stauffer S, Ford B. Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. J Pathol Inform. 2015;6.

[73] Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inf Assoc. 2018;25.

[74] Shivade C, Malewadkar P, Fosler-Lussier E, Lai AM. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. J Biomed Inform. 2015;58.

[75] Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. Appl Clin Inform. 2016;7.

[76] Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Informatics Assoc. 2015;22.

[77] Sun H, Depraetere K, Roo J, Mels G, Vloed B, Twagirumukiza M. Semantic processing of EHR data for clinical research. J Biomed Inform. 2015;58.

[78] Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inf. 2017;73.

[79] Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the Patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inf. 2017;26.

[80] Jovanovic J, Bagheri E, Jovanović J. Semantic annotation in biomedicine: the current landscape. J Biomed Semant. 2017;8.

[81] UK EQUATOR Centre. The EQUATOR Network;.

[82] Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Informatics Assoc. 2016;23.

[83] Vuokko R, Makela-Bengs P, Hypponen H, Lindqvist M, Doupi P, Mäkelä-Bengs P. Impacts of structuring the electronic health record: results of a systematic literature review from the perspective of secondary use of patient data. Int J Med Inform. 2017;97.

[84] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. TRIPOD Group Circ. 2015;131.

[85] Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61.

[86] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015 Oct;12(10):e1001885.

[87] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting

diagnostic accuracy studies. BMJ. 2015 Oct;p. h5527.

[88] Moher D, Liberati A, Tetzlaff J, and DGA. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009 Jul;339(jul21 1):b2535–b2535.

[89] The EndNote Team. EndNote X9; 2013.

[90] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews. 2016 Dec;5(1).

[91] Veritas Health Innovation. Covidence systematic review software; 2020.

[92] Matentzoglu N, Malone J, Mungall C, Stevens R. MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semantics. 2018;9.

[93] Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta. 2013;760.

[94] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol. 2019;10.

[95] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45.

[96] Afshar M, Dligach D, Sharma B, Cai X, Boyda J, Birch S. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. J Am Med Inform Assoc. 2019;26.

[97] Alnazzawi N, Thompson P, Ananiadou S. Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource. PLoS One. 2016;11.

[98] Atutxa A, Perez A, Casillas A. Machine Learning Approaches on Diagnostic Term Encoding with the ICD for Clinical Documentation. IEEE J Biomed Heal Informatics. 2018;22.

[99] Barrett N, Weber-Jahnke JH, Thai V. Engineering natural language processing solutions for structured information from clinical text: extracting sentinel events from palliative care consult letters. Stud Health Technol Inform. 2013;192.

[100] Becker M, Bockmann B. Extraction of UMLS(R) Concepts Using Apache cTAKES for German Language. Stud Health Technol Inform. 2016;223.

[101] Becker M, Kasper S, Böckmann B, Jöckel KH, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. Int J Med Inform. 2019;127.

[102] Bejan CA, Wei WQ, Denny JC. Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text. J Am Med Informatics Assoc. 2015;22.

[103] Castro E, Iglesias A, Martínez P, Castaño L. Automatic Identification of Biomedical Concepts in Spanish-language Unstructured Clinical Texts. German Research Cent for Artificial, Intelligence - DFKI GmbH, Kaiserslautern, Germany Seattle, WA, USA: ACM; 2010.

[104] Catling F, Spithourakis GP, Riedel S. Towards automated clinical coding. Int J Med Inform. 2018;120.

[105] Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Medinfo. 2004;11.

[106] Chen J, Zheng J, Yu H. Finding Important Terms for Patients in Their Electronic Health Records: A Learning-to-Rank Approach Using Expert Annotations. JMIR Med informatics. 2016;4.

[107] Chiaramello E, Pinciroli F, Bonalumi A, Caroli A, Tognola G. Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. J Biomed Inform. 2016;63.

[108] Chodey KP, Hu G. In: Clinical text analysis using machine learning methods; 2016. .

[109] Chung J, Murphy S. Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. AMIA Annu Symp Proc. 2005;p. 131–135.

[110] Combi C, Zorzi M, Pozzani G, Moretti U, Arzenton E. From narrative descriptions to MedDRA: automagically encoding adverse drug reactions. J Biomed Inform. 2018;84.

[111] Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: The state of the art at i2b2 2010. J Am Med Informatics Assoc. 2011;18.

[112] Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet Med. 2019;21.

[113] Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: An evaluation of a new Java implementation of MetaMap. J Am Med Informatics Assoc. 2017;24.

[114] Divita G, Zeng QT, Gundlapalli AV, Duvall S, Nebeker J, Samore MH. Sophia: A Expedient UMLS Concept Extraction Annotator. AMIA Annu Symp Proc. 2014;2014.

[115] Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. J Biomed Inform. 2018;80.

[116] Falis M, Pajak M, Lisowska A, Schrempf P, Deckers L, Mikhael S, et al. Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Association for Computational Linguistics; 2019. .

[117] Ferrao JC, Janela F, Oliveira MD, Martins HMG. Using Structured EHR Data and SVM to Support ICD-9-CM Coding. In: 2013 IEEE International Conference on Healthcare Informatics. IEEE; 2013. p. 511–516.

[118] Gerbier S, Yarovaya O, Gicquel Q, Millet AL, Smaldore V, Pagliaroli V, et al. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. BMC Medical Informatics and Decision Making. 2011 Jul;11(1).

[119] Goicoechea Salazar JA, Nieto García MA, Laguna Téllez A, Canto Casasola VD, Rodríguez Herrera J, Murillo CF. Development of an automated coding system to retrieve and analyze diagnostic information stored in hospital emergency department records. Emergencias. 2013;25.

[120] Hamid H, Fodeh SJ, Lizama AG, Czlapinski R, Pugh MJ, LaFrance WC. Validating a natural language processing tool to exclude psychogenic nonepileptic seizures in electronic medical record-based epilepsy research. Epilepsy Behav. 2013;29.

[121] Hassanzadeh H, Nguyen A, Koopman B. Evaluation of Medical Concept Annotation Systems on Clinical Records; 2016.

[122] Helwe C, Elbassuoni S, Geha M, Hitti E, Obermeyer CM. CCS Coding of Discharge Diagnoses via Deep Neural Networks. In: Proceedings of the 2017 International Conference on Digital Health. ACM; 2017. .

[123] Hersh W, Mailhot M, Arnott-Smith C, Lowe H. Selective automated indexing of findings and diagnoses in radiology reports. J Biomed Inform. 2001;34.

[124] Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med. 2015;69.

[125] Jindal P, Roth D. Extraction of events and temporal expressions from clinical narratives. J Biomed Inform. 2013;46.

[126] Kang BY, Kim DW, Kim HG. Two-phase chief complaint mapping to the UMLS metathesaurus in Korean Electronic Medical Records. IEEE Trans Inf Technol Biomed. 2009;13.

[127] Kersloot MGMG, Lau F, Abu-Hanna A, Arts DLDL, Cornet R. Automated SNOMED CT concept and attribute relationship detection through a web-based implementation of cTAKES. J Biomed Semantics. 2019;10.

[128] König M, Sander A, Demuth I, Diekmann D, Steinhagen-Thiessen E. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. PLoS One. 2019;14.

[129] Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy

detection. BMC Medical Informatics and Decision Making. 2015 May;15(1).

[130] Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med informatics. 2019;7.

[131] Liu C, Ta CN, Rogers JR, Li Z, Lee J, Butler AM. Ensembles of natural language processing systems for portable phenotyping solutions. J Biomed Inform. 2019;100.

[132] Lowe HJ, Huang Y, Regula DP. Using a statistical natural language Parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports. AMIA Annu Symp Proc. 2009;2009.

[133] Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Informatics Assoc. 2014;21.

[134] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform. 2006;39.

[135] Meystre SM, Thibault J, Shen S, Hurdle JF, South BR. Automatically detecting medications and the reason for their prescription in clinical narrative text documents. Stud Health Technol Inform. 2010;160.

[136] Minard AL, Ligozat AL, Abacha AB, Bernhard D, Cartoni B, Deléger L. Hybrid methods for improving information access in clinical documents: Concept, assertion, and relation identification. J Am Med Informatics Assoc. 2011;18.

[137] Mishra R, Burke A, Gitman B, Verma P, Engelstad M, Haendel MA. Data-driven method to enhance craniofacial and oral phenotype vocabularies. J Am Dent Assoc. 2019;150.

[138] Nguyen AN, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J. Computer-assisted diagnostic coding: effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. AMIA Annu Symp proceedings AMIA Symp. 2018;2018.

[139] Oellrich A, Collier N, Smedley D, Groza T. Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. PLoS One. 2015;10.

[140] Patrick JD, Nguyen DHM, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Informatics Assoc. 2011;18.

[141] Pérez A, Atutxa A, Casillas A, Gojenola K, Sellart Á. Inferred joint multigram models for medical term normalization according to ICD. Int J Med Inform. 2018;110.

[142] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018;18.

[143] Roberts K, Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. J Am Med Informatics Assoc. 2011;18.

[144] Rousseau JF, Ip IK, Raja AS, Valtchinov VI, Cochon L, Schuur JD. Can automated retrieval of data from emergency department physician notes enhance the imaging order entry process? Appl Clin Inform. 2019;10.

[145] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Informatics Assoc. 2010;17.

[146] Shoenbill K, Song Y, Gress L, Johnson H, Smith M, Mendonca EA. Natural language processing of lifestyle modification documentation. Health Informatics Journal. 2019 Feb;26(1):388–405.

[147] Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: An open source medication extraction and normalization tool for clinical text. J Am Med Informatics Assoc. 2014;21.

[148] Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, et al. Building an automated problem list based on natural language processing: lessons learned in the early phase of development. AMIA Annu Symp Proc. 2008 Nov;p. 687–691.

[149] Soriano IM, Pena JLC, Breis JTF, Roman IS, Barriuso AA, Baraza DG. Snomed2Vec: Representation of SNOMED CT Terms with Word2Vec.

In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2019. .

[150] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Informatics Assoc. 2018;25.

[151] Spasić I, Zhao B, Jones CB, Button K. KneeTex: An ontology-driven system for information extraction from MRI reports. J Biomed Semantics. 2015;6.

[152] Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Informatics Assoc. 2013;20.

[153] Sung SF, Chen K, Wu DP, Hung LC, Su YH, Hu YH. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study. Int J Med Inform. 2018;112.

[154] Tchechmedjiev A, Abdaoui A, Emonet V, Zevio S, Jonquet C. SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. BMC Bioinformatics. 2018;19.

[155] Ternois I, Escudie JB, Benamouzig R, Duclos C. Development of an automatic coding system for digestive endoscopies. Stud Health Technol Inform. 2018;255.

[156] Travers DA, Haas SW. Evaluation of Emergency Medical Text Processor, a system for cleaning chief complaint text data. Acad Emerg Med. 2004;11.

[157] Tulkens S, Šuster S, Daelemans W. Unsupervised concept extraction from clinical text through semantic composition. J Biomed Inform. 2019;91.

[158] Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for Pharmacovigilance: natural language processing analysis in Japanese. JMIR Med informatics. 2018;6.

[159] Valtchinov VI, Lacson R, Wang A, Khorasani R. Comparing Artificial Intelligence Approaches to Retrieve Clinical Reports Documenting Implantable Devices Posing MRI Safety Risks. J Am Coll Radiol. 2019;S1546–1440.

[160] Wadia R, Akgun K, Brandt C, Fenton BT, Levin W, Marple AH. Comparison of natural language processing and manual coding for the identification of cross-sectional imaging reports suspicious for lung Cancer. JCO Clin cancer informatics. 2018;2.

[161] Walker G, Soysal E, Xu H. Development of a natural language processing tool to extract radiation treatment sites. Cureus. 2019;11.

[162] Xie X, Xiong Y, Yu PS, Zhu Y. EHR Coding with Multi-scale Feature Attention and Structured Knowledge Graph Propagation. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. ACM; 2019. .

[163] Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc. 2011;2011.

[164] Yadav K, Sarioglu E, Smith M, Choi HA. Automated outcome classification of emergency department computed tomography imaging reports. Acad Emerg Med. 2013;20.

[165] Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak. 2019;19.

[166] Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018;19.

[167] Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. J Biomed Inform. 2013;46.

[168] Zhou X, Han H, Chankai I, Prestrud A, Brooks A. Approaches to text mining for clinical medical records. In: Proceedings of the 2006 ACM symposium on Applied computing - SAC '06. ACM Press; 2006. .

[169] Zhou L, Lu Y, Vitale CJ, Mar PL, Chang F, Dhopeshwarkar N. Representation of information about family relatives as structured data

in electronic health records. Appl Clin Inform. 2014;5.

[170] Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc. 2011;2011.

[171] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008;p. 128–144.

[172] Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW. How many medication orders are entered through free-text in EHRs?–a study on hypoglycemic agents. AMIA Annu Symp Proc AMIA Sym. 2012;2012.

[173] Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Washington, DC). 2013;1.

[174] Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K. Towards a semantic lexicon for clinical natural language processing. AMIA Ann Symp Proc AMIA Symp. 2012;2012.

[175] Szlosek DA, Ferrett J. Using machine learning and natural language processing algorithms to Automate the evaluation of clinical decision support in electronic medical record systems. EGEMS (Washington, DC). 2016;4.

[176] Ruch P, Baud R, Geissbuhler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif Intell Med. 2003;29.

[177] Salmasian H, Freedberg DE, Friedman C. Deriving comorbidities from medical records using natural language processing. J Am Med Inform Assoc. 2013;20.

[178] Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care 2014; 2014.

[179] Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. 2015;15.

[180] Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol. 2014;179.

[181] Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD. Web-based real-time case finding for the population health Management of Patients with Diabetes Mellitus: a prospective validation of the natural language processing-based algorithm with statewide electronic medical records. JMIR Med Inform. 2016;4.

[182] U S National Library of Medicine. RxNorm; 2014.

[183] Masanz J, Pakhomov SV, Xu H, Wu ST, Chute CG, Liu H. Open source clinical NLP - more than any single system. AMIA Jt Summits Transl Sci Proc. 2014;2014.

[184] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18.

[185] National Library of Medicine (US). Introduction to the UMLS. In: UMLS® Reference Manual; 2009. .

[186] Choi JD, Palmer M. Guidelines for the clear style constituent to dependency conversion. vol. 12; 2012.

[187] Oliver I. Programming classics: implementing the world's best algorithms: Prentice Hall; 1993.

[188] PHP Group. similar_text. PHP Group;.

[189] SNOMED International. SNOMED CT Machine Readable Concept Model; 2017.

[190] Finan S. Dictionary Creator GUI; 2017.

[191] Dror R, Baumer G, Shlomov S, Reichart R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2018. .

[192] Kodra Y, de la Paz MP, Coi A, Santoro M, Bianchi F, Ahmed F, et al. Data Quality in Rare Diseases Registries. In: Advances in Experimental Medicine and Biology. Springer International Publishing; 2017. p. 149–164.

[193] Kodra Y, Weinbach J, de-la Paz MP, 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 Aug;15(8):1644.

[194] 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 Journal of Rare Diseases. 2019 Jul;14(1).

[195] Rubinstein YR, Robinson PN, Gahl WA, Avillach P, Baynam G, Cederroth H, et al. The case for open science: rare diseases. JAMIA Open. 2020 Sep;3(3):472–486.

[196] van Soest Johan, Chang S, Ole M, Marco P, van den Berg Bob, Alexander M, et al. Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data. Studies in Health Technology and Informatics. 2018;247(Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth):581–585.

[197] 8. IRDiRC - Inserm EJP RD, International Rare Diseases Research Consortium; 2020.

[198] Hooft R, Goble C, Evelo C, Roos M, Sansone S, Ehrhart F, et al. ELIXIR-EXCELERATE D5.3: Bring Your Own Data (BYOD). 2019;.

[199] Roos M, Lopes P. Bring your own data parties and beyond: make your data linkable to speed up rare disease research. Rare Diseases and Orphan Drugs. 2014;p. 21–24.

[200] Roos M, Gray AJ, Waagmeester A, Thompson M, Kaliyaperumal R, Van Der Horst E, et al. Bring Your Own Data Workshops: A Mechanism to Aid Data Owners to Comply with Linked Data Best Practices. In: SWAT4LS; 2014. .

[201] The Human Phenotype Ontology; 2020.

[202] Orphanet Rare Disease Ontology (ORDO); 2020.

[203] European Platform on Rare Disease Registration. Set of Common Data Elements for Rare Disease Registration; 2019.

[204] European Commission. European Rare Disease Registry Infrastructure (ERDRI); 2020.

[205] European Commission. European Reference Networks (ERNs); 2020.

[206] European Commission. Call for project proposals under the Annual Work Programme 2019, 3rd EU Health Programme; 2019.

[207] ERN on Rare Multisystemic Vascular Diseases (VASCERN); 2020.

[208] Vascular Anomalies working group (VASCA); 2020.

[209] W3C. Data Catalog Vocabulary (DCAT) - Version 2; 2020.

[210] FAIR Data Team. FAIR Data Point design specification; 2019.

[211] Kersloot MG, Jacobsen A, Groenen K, dos Santos Vieira B, Kaliyaperumal R, Abu-Hanna A, et al.. The Joint Research Council's Common Data Elements and their implementations on an electronic Case Report Form; 2021.

[212] VASCA Common Data Elements (CDE) - Datasets; 2020.

[213] The iCRF Generator; 2020.

[214] LUMC. Semantic data model of the set of common data elements for rare disease registration; 2020.

[215] Research Data Alliance FAIR Data Maturity Model Working Group. FAIR Data Maturity Model: specification and guidelines. 2020;.

[216] Wilkinson MD, Dumontier M, Sansone SA, Bonino da Silva Santos LO, Prieto M, Batista D, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific Data. 2019;6(1):174.

[217] Bonino da Silva Santos LO, Wilkinson M, Kuzniar A, Kaliyaperumal R, Thompson M, Dumontier M, et al. FAIR Data Points Supporting Big Data Interoperability. In: Enterprise Interoperability in the Digitized and Networked Factory of the Future. ISTE Press; 2016. .

[218] W3C. RDF 1.1 Concepts and Abstract Syntax; 2014.

[219] European Commission. European Directory of Registries (ERDRI.dor); 2020.

[220] European Commission. ERDRI Metadata Repository (ERDIR.mdr); 2020.

[221] European Joint Programme on Rare Diseases. Semantic data model of the set of common data elements for rare disease registration; 2020.

[222] Austrian Institute of Technology GmbH. EU-PID - European Patient Identity Management; 2020.

[223] European Commission. European Rare Disease Registry Infrastructure (ERDRI); 2020.

[224] Gene42 Inc. Phenotips; 2020.

[225] EMBL-EBI. Zooma ontology annotation; 2020.

[226] System for Ontology-based Re-coding and Technical Annotation (SORTA); 2020.

[227] Pang C, Sollie A, Sijtsma A, Hendriksen D, Charbon B, de Haan M, et al. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database. 2015 1;2015.

[228] U S National Library of Medicine. MEDLINE/PubMed Baseline Repository (MBR); 2019.

[229] Open Knowledge Foundation. Open Data Handbook - Machine Readable; 2018.

[230] European Commission. Guidelines on Data Management in Horizon 2020; 2013. December.

[231] European Commission. Rare diseases | Public Health; 2016.

[232] Lochmüller H, TorrentFarnell J, Le Cam Y, Jonker AH, Lau LP, Baynam G, et al. The International Rare Diseases Research Consortium: Policies and Guidelines to maximize impact. European Journal of Human Genetics. 2017;25(12):1293–1302.

[233] Castor EDC. Castor EDC API Documentation; 2021.

[234] Castor EDC FAIR Data Point - Registry of Vascular Anomalies; 2021.

[235] FAIR Data Team. FAIR Data Point Home; 2021.

[236] Dublin Core Metadata Initiative Usage Board. DCMI: DCMI Metadata Terms; 2020.

[237] Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017;7:18647.

[238] Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL, et al. RightField: embedding ontology annotation in spreadsheets. Bioinformatics. 2011 7;27(14):2021–2022.

[239] Sernadela P, González-Castro L, Carta C, van der Horst E, Lopes P, Kaliyaperumal R, et al. Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer. BioMed Research International. 2017;2017:8327980.

[240] Potencier F. Twig - The flexible, fast, and secure PHP template engine; 2020.

[241] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association. 2012;19(1):54–60.

[242] Clinical Data Interchange Standards Consortium. ODM-XML; 2013.

[243] National Cancer Institute. Biomedical Research Integrated Domain Group (BRIDG) Model; 2019.

[244] National Library of Medicine. NIH Common Data Elements (CDE) Repository; 2021.

[245] Wassef M, Blei F, Adams D, Alomari A, Baselga E, Berenstein A, et al. Vascular anomalies classification: recommendations from the International Society for the Study of Vascular Anomalies. Pediatrics. 2015;136(1):e203–e214.

[246] International Society for the Study of Vascular Anomalies. About ISSVA; 2020.

[247] Mulliken JB, Glowacki J. Hemangiomas and vascular malformations in infants and children: a classification based on endothelial characteristics. Plastic and reconstructive surgery. 1982;69(3):412–422.

[248] International Society for the Study of Vascular Anomalies. ISSVA Classification of Vascular Anomalies; 2018.

[249] Orphanet. Procedural document: Orphanet nomenclature and classification of rare diseases; 2020.

[250] Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI thesaurus. Comparative and functional genomics. 2004;5(8):648–654.

[251] Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207–D1217.

[252] HUGO Gene Nomenclature Committee at the European Bioinformatics Institute. HUGO Gene Nomenclature Committee; 2021.

[253] Preston-Werner T. Semantic Versioning 2.0.0; 2013.

[254] van Damme P, Kersloot MG, Dos Santos Viera B, Schultze Kool L, Cornet R. The International Society for the Study of Vascular Anomalies Ontology; 2021.

[255] van Damme P, Kersloot MG, Dos Santos Viera B, Schultze Kool L, Cornet R. The International Society for the Study of Vascular Anomalies Ontology; 2021.

[256] Schober D, Kusnierczyk W, Lewis SE, Lomax J, et al. Towards naming conventions for use in controlled vocabulary and ontology engineering. In: The 10th Annual Bio-Ontologies Meeting; 2007. .

[257] The OBO Foundry. OBO Foundry Identifier Policy; 2020.

[258] Kamdar MR, Tudorache T, Musen MA. A systematic analysis of term reuse and term overlap across biomedical ontologies. Semant Web. 2017;8(6):853–871.

[259] Faria D, Pesquita C, Santos E, Cruz IF, Couto FM. AgreementMakerLight 2.0: Towards efficient large-scale ontology matching. In: CEUR Workshop Proc.. vol. 1272; 2014. p. 457–460.

[260] Jiménez-ruiz E, Grau BC, Zhou Y. LogMap 2.0: towards logic-based, scalable and interactive ontology matching Ernesto. In: SWAT4LS; 2011. p. 2–3.

[261] Zhao M, Zhang S, Li W, Chen G. Matching biomedical ontologies based on formal concept analysis. J Biomed Semantics. 2018;9(1):1–27.

[262] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research. 2011;39(suppl_2):W541–W545.

[263] Kersloot MG, Jacobsen A, Groenen KHJ, dos Santos Vieira B, Kaliyaperumal R, Abu-Hanna A, et al. De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data. 2021 Oct;122:103897.

[264] Miller AR, Tucker C. Health Information Exchange, System Size and Information Silos. SSRN Electron J. 2011;.

[265] Griggs RC, Batshaw M, Dunkle M, Gopal-Srivastava R, Kaye E, Krischer J, et al. Clinical research for rare disease: Opportunities, challenges, and solutions. Mol Genet Metab. 2009;96(1):20–26.

[266] Cruz IF, Xiao H. The role of ontologies in data integration. Eng Intell Syst. 2005;13(4):245–252.

[267] European Joint Programme on Rare Diseases. Coordinated Access to Data & Resources; 2021.

[268] Harrow I, Jiménez-Ruiz E, Splendiani A, Romacker M, Woollard P, Markel S, et al. Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. J Biomed Semantics. 2017;8(1):1–13.

[269] EOSCpilot. EOSCpilot framework of FAIR data stewardship skills for science and scholarship, and draft recommendations on FAIR training; 2018.

[270] Koers H, Bangert D, Hermans E, van Horik R, de Jong M, Mokrane M. Recommendations for Services in a FAIR Data Ecosystem. Patterns. 2020 Aug;1(5):100058.

[271] Devaraju A, Huber R. An automated solution for measuring the progress toward FAIR research data. Patterns. 2021 nov;2(11):100370.

[272] GO FAIR. Data Together COVID-19 Appeal and Actions; 2020.

[273] Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, et al. The Availability of Research Data Declines Rapidly with Article Age. Current Biology. 2014 Jan;24(1):94–97.

[274] The Lancet. The REWARD Statement; 2014.

[275] Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. JAMA. 2013 Apr;309(13):1351.

[276] Todd OM, Burton JK, Dodds RM, Hollinghurst J, Lyons RA, Quinn TJ, et al. New Horizons in the use of routine data for ageing research. Age and Ageing. 2020 Feb;49(5):716–722.

[277] Ross JS, Waldstreicher J, Bamford S, Berlin JA, Childers K, Desai NR, et al. Overview and experience of the YODA Project with clinical trial data sharing after 5 years. 2018 Nov;5(1).

[278] Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing Clinical Trial Data: A Proposal from the International Committee of Medical Journal Editors. 2016 Jan;13(1):e1001950.

[279] Chawinga WD, Zinn S. Global perspectives of research data sharing: A systematic literature review. 2019 Apr;41(2):109–122.

[280] Demotes-Mainard J, Cornu C, Guérin A, Bertoye PH, Boidin R, Bureau S, et al. How the new European data protection regulation affects clinical research and recommendations? Therapies. 2019 Feb;74(1):31–42.

[281] Naudet F, Siebert M, Pellen C, Gaba J, Axfors C, Cristea I, et al. Medical journal requirements for clinical trial data sharing: Ripe for improvement. 2021 Oct;18(10):e1003844.

[282] Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials. BMJ. 2017 Jun;p. j2372.

[283] Mons B. FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services. Data Intelligence. 2019 Mar;1(1):22–42.

[284] Miron L, Gonçalves RS, Musen MA. Obstacles to the reuse of study metadata in ClinicalTrials.gov. Scientific Data. 2020 Dec;7(1).

[285] Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. Journal of Biomedical Informatics. 2020 Jul;107:103421.

[286] Clinical Data Interchange Standards Consortium. CDASH; 2021.

[287] Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database. 2016;2016:bav121.

[288] Teare HJA, Prictor M, Kaye J. Reflections on dynamic consent in biomedical research: the story so far. European Journal of Human Genetics. 2020 Nov;29(4):649–656.

[289] Wilkinson T, Sinha S, Peek N, Geifman N. Clinical trial data reuse – overcoming complexities in trial design and data sharing. Trials. 2019 Aug;20(1).

[290] Dron L, Golchi S, Hsu G, Thorlund K. Minimizing control group allocation in randomized trials using dynamic borrowing of external control data – An application to second line therapy for non-small cell lung cancer. Contemporary Clinical Trials Communications. 2019 Dec;16:100446.

[291] Deist TM, Dankers FJWM, Ojha P, Marshall MS, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiotherapy and Oncology. 2020 Mar;144:189–200.

Summary

Data are the foundation of modern medicine: they contribute to building evidence that is incorporated into the body of scientific knowledge and clinical practice. To gather new evidence, one sets up a research project in which data are collected, analyzed, and contextualized. Throughout such a project, proper management and stewardship of data are essential. The Research Data Life Cycle (Figure 1) guides researchers through the process of setting up and conducting research in line with these practices by describing all steps taken in a research project, as well as the associated data management tasks.

The FAIR Principles, similarly, are designed to guide the implementation of good data management and stewardship. They state that research data and metadata should be Findable, Accessible, Interoperable, and Reusable, both for humans and machines. Workflows have been developed over the years to make data and metadata FAIR step by step. However, present workflows are designed to be executed after research projects have been already conducted and data are collected, rather than throughout the life cycle of a research project. In this thesis we, therefore, aimed to incorporate the FAIRification steps (i.e., steps to make data and metadata more Findable, Accessible, Interoperable, and Reusable) into the Research Data Life Cycle, ensuring that data are FAIRified throughout the research process rather than after the project is completed. To investigate this, we distinguish three strands of research, which are addressed in the three parts of this thesis. These parts are summarized below.

Part I. State of FAIR

The FAIR Data Principles are rapidly being adopted by many research institutes and funders worldwide. However, little is known about the knowledge, perceptions, and efforts of individual clinical researchers and research support staff regarding data FAIRification.

In Chapter 2 we, therefore, assessed the awareness and attitudes of clinical researchers and research support staff regarding data FAIRification. We distributed a questionnaire to researchers and support staff in six Dutch University Medical Centers and Electronic Data Capture platform users. We found that 62.8% of the researchers and 81.0% of the support staff are currently undertaking at least some effort to achieve *any* aspect of FAIR (i.e., Findability, Accessibility, Interoperability, or Reusability). 11.0% and 23.8%, respectively, address all of the FAIR aspects. 94.7% of the researchers are aware that their data being FAIR might be useful for others and 89.3% are, given the right resources and support, willing to FAIRify their data.

In Chapter 3 we used the data of this questionnaire to gain insight into clinical researchers' understanding of the FAIR Principles and their experience with data FAIRification. We found that most researchers were unaware of the Principles' emphasis on both human and machine readability, as their FAIRification efforts were primarily focused on achieving human readability (93.9%), rather than machine readability (31.2%).

Part II. NLP and FAIR

Much of the data present in Electronic Health Records (EHRs) are stored as unstructured free text, as clinicians often resort to making free-text notes. These notes may be useful

for clinical research, but cannot be readily interpreted by a machine. Natural Language Processing (NLP) algorithms can make free text machine-interpretable by attaching it to ontology concepts, formal specifications of concepts and relations in a domain.

In Chapter 4 we reviewed the current methods used for developing and evaluating NLP algorithms that map clinical text fragments onto ontology concepts. We found many heterogeneous approaches to reporting on the development and evaluation of these algorithms. In order to standardize the evaluation of algorithms and reduce heterogeneity between studies, we developed a list of sixteen recommendations for future studies that can be used along with adherence to a generic reporting standard.

In Chapter 5 we developed and evaluated an NLP application with generic algorithms for the detection of (misspelled) concepts and of relationships between them. We evaluated the application by encoding free-text oncology charts. The evaluation results show that our application can detect oncology concepts and relationships with high precision. These concepts and relationships can be used to encode clinical narratives, and can thus substantially reduce manual chart abstraction efforts, saving time for clinicians and researchers.

Part III. FAIR by design

Existing FAIRification workflows are usually carried out *post hoc*: after the research project is conducted and data are collected, rather than throughout the life cycle of a research project. Researchers would benefit from a *de-novo* approach, where data are made FAIR automatically and in real-time, upon collection.

In Chapter 6 we present a workflow for *de-novo* FAIRification: a workflow where FAIRification steps are incorporated in the process of setting up and collecting data for a registry or research project. This facilitates automated, real-time FAIRification, without any intervention from data management and data entry personnel. The workflow is applied to a registry for vascular anomalies, and can, to a large extent, be reused by other research projects.

In Chapter 7 we describe the implementation and evaluation of Chapter 6's workflow in an Electronic Data Capture (EDC) system, the place where medical research data are often collected and stored via electronic Case Report Forms (eCRFs). Our implementation ensures that eCRF data entered into an EDC system can be transformed into machine-readable, FAIR data using a semantic data model (a canonical representation of the data, based on ontology concepts and semantic web standards) and mappings from the model to questions on the eCRF. Its application in the registry for vascular anomalies and the evaluation results show that the method can be used to make clinical research data FAIR when they are entered in an eCRF, without any intervention from data management and data entry personnel.

In Chapter 8 we describe the process of transforming the International Society for the Study of Vascular Anomalies (ISSVA) classification, a classification for vascular anomalies that enables specialists to unambiguously classify diagnoses, into an ontology. The classification was only available in a human-readable format, but is now made machine-readable and interoperable, and now includes mappings to existing ontologies. We be-

lieve that the ontology will increase the interoperability between clinical studies and registries and, therefore, will contribute to FAIRer vascular anomaly research.

Conclusion

Chapter 9 presents an overall discussion of the work in this thesis. The conclusions drawn in this thesis are summarized in Figure 1.

The FAIR Principles are gaining more and more traction. However, many researchers are currently unaware of the (meaning of the) FAIR Principles or how to apply them to their work. The work presented in this thesis quantifies the lack of FAIR awareness of clinical researchers and research support staff and provides a FAIRification workflow with tooling that is incorporated into the current way in which researchers work, i.e., the Research Data Life Cycle. Workflows and tools contribute to making data FAIRification more scalable. User-friendly tools that are integrated with the software used by researchers play a crucial role in this, since they facilitate automated data FAIRification throughout the course of a research project and take the majority of the work off researchers' shoulders. To accelerate FAIRification at scale, policy makers, funders and research institutes need to work together to provide standards, methods, support, tools, and funding to the research community, effectively making FAIRification of research data a joint mission. Ultimately, this paves the way for a future where collected data brings maximal value to patient care.

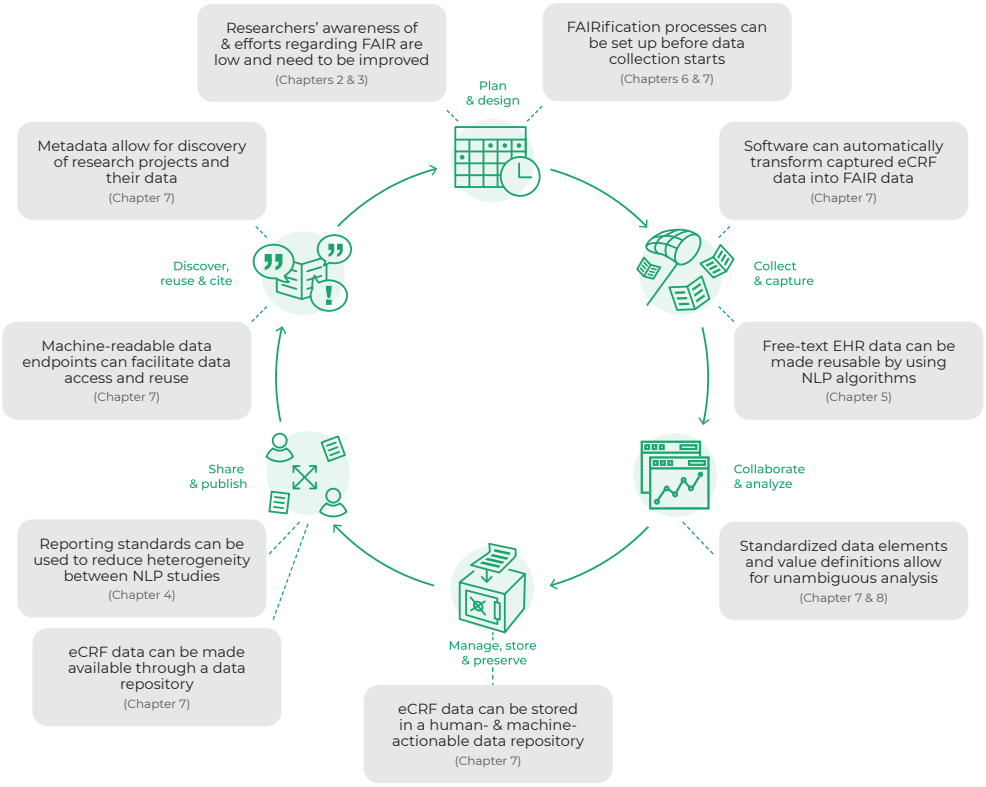


Figure 1: Summary of main findings and conclusions, mapped to the Research Data Life Cycle

FAIR: Findable, Accessible, Interoperable and Reusable, NLP: Natural Language Processing, eCRF: electronic Case Report Form, EHR: Electronic Health Record

Samenvatting

Data liggen aan de basis van de hedendaagse geneeskunde: ze dragen bij aan het verkrijgen van bewijs dat vervolgens wordt opgenomen in de wetenschappelijke literatuur en de klinische praktijk. Om nieuw bewijs te verzamelen, voert men in het algemeen een onderzoeksproject uit waarin data worden verzameld, geanalyseerd en gecontextualiseerd. Tijdens een dergelijk project zijn *data management* en *data stewardship* essentieel. De Research Data Life Cycle (Figuur 2) leidt onderzoekers door het proces van het opzetten en uitvoeren van onderzoek. De Life Cycle beschrijft alle stappen die in een onderzoeksproject worden genomen met de bijbehorende data management-taken.

Ook de FAIR Principles zijn opgesteld ten behoeve van goed data management en data stewardship. Ze stellen dat onderzoeksdata Findable, Accessible, Interoperable en Reusable (vindbaar, toegankelijk, interoperabel en herbruikbaar) moeten zijn, zowel voor mensen als voor machines. Door de jaren heen zijn er workflows ontwikkeld om data en metadata stap voor stap FAIR te maken. De huidige workflows zijn echter ontworpen om te worden uitgevoerd nadat onderzoeksprojecten al zijn uitgevoerd en data zijn verzameld, in plaats van gedurende het verloop van een onderzoeksproject. Het in dit proefschrift beschreven onderzoek was erop gericht de FAIRificatie-stappen (de stappen om data en metadata meer Findable, Accessible, Interoperable, and Reusable te maken) op te kunnen nemen in de Research Data Life Cycle. Dit om ervoor te zorgen dat data FAIR worden gemaakt gedurende het onderzoeksproces, in plaats van na het voltooien van het project. We onderscheiden hierbij drie onderzoekslijnen, die in de drie delen van dit proefschrift aan de orde komen. Deze delen zijn hieronder samengevat.

Deel I. De huidige staat van FAIR

De FAIR Data Principles zijn snel overgenomen door veel onderzoeksinstituten en subsidieverstrekkers over de hele wereld. Er is echter weinig bekend over de kennis, percepties en inspanningen van individuele klinische onderzoekers en onderzoeksondersteuners (zoals data managers en data stewards) met betrekking tot data FAIRificatie.

In Hoofdstuk 2 hebben we daarom het bewustzijn en de houding van klinische onderzoekers en onderzoeksondersteuners ten aanzien van data FAIRificatie beoordeeld. We hebben hiertoe een vragenlijst verspreid onder onderzoekers en ondersteunend personeel in zes Nederlandse Universitair Medische Centra en gebruikers van een Electronic Data Capture-platform. We ontdekten dat 62,8% van de onderzoekers en 81,0% van het ondersteunend personeel momenteel op zijn minst enige moeite doen om *enig* aspect van FAIR te bereiken (d.w.z. Findability, Accessibility, Interoperability of Reusability). Respectievelijk 11,0% en 23,8% behandelen alle FAIR-aspecten. 94,7% van de onderzoekers is zich ervan bewust dat het FAIR zijn van onderzoeksdata nuttig kan zijn voor anderen en 89,3% is, met de juiste middelen en ondersteuning, bereid om onderzoeksdata FAIR te maken.

In Hoofdstuk 3 hebben we de data van de vragenlijst gebruikt om inzicht te krijgen in het begrip dat klinische onderzoekers hebben van de FAIR Principles en hun ervaring met data FAIRificatie. De meeste onderzoekers bleken zich niet bewust te zijn van de nadruk die de Principles leggen op leesbaarheid van metadata en data voor zowel mensen als machines, aangezien hun FAIRificatie-inspanningen voornamelijk gericht waren

op het realiseren van leesbaarheid voor mensen (93,9%), in plaats van leesbaarheid voor machines (31,2%).

Deel II. NLP en FAIR

Veel van de data in Elektronisch Patiënten Dossiers (EPDs) worden opgeslagen als ongestructureerde vrije tekst, omdat clinici vaak hun bevindingen noteren in notities. Deze notities kunnen nuttig zijn voor klinisch onderzoek, maar kunnen niet gemakkelijk worden geïnterpreteerd door een machine. Natural Language Processing (NLP) algoritmen kunnen vrije tekst interpreteerbaar maken voor machines door tekstfragmenten te koppelen aan concepten in een ontologie, dat wil zeggen formele specificaties van concepten en relaties in een domein.

In Hoofdstuk 4 hebben we de huidige methoden besproken die worden gebruikt voor het ontwikkelen en evalueren van NLP-algoritmen die klinische tekstfragmenten koppelen aan ontologieconcepten. We hebben een hoge heterogeniteit geconstateerd in de wijze van rapportage van de ontwikkeling en evaluatie van deze algoritmen. Om de evaluatie van algoritmen te standaardiseren en de heterogeniteit tussen onderzoeken te verminderen, hebben we een lijst met zestien aanbevelingen voor toekomstige onderzoeken ontwikkeld die kunnen worden gebruikt in combinatie met het gebruik van een generieke rapportagestandaard.

In Hoofdstuk 5 hebben we een NLP-toepassing ontwikkeld en geëvalueerd met generieke algoritmen voor de detectie van concepten en hun onderlinge relaties in vrije tekst met mogelijk spelfouten. We hebben de toepassing geëvalueerd door voortgangsnotities van oncologiepatiënten te annoteren. De evaluatieresultaten laten zien dat onze applicatie oncologische concepten en relaties met hoge precisie kan detecteren. Deze concepten en relaties kunnen gebruikt worden voor het annoteren van klinische notities, en kunnen dus de noodzaak voor handmatige screening van dossiers aanzienlijk verminderen, wat tijd bespaart voor clinici en onderzoekers.

Deel III. “FAIR by design”

Bestaande FAIRificatie-workflows worden meestal *post hoc* uitgevoerd: nadat het onderzoeksproject is uitgevoerd en de data zijn verzameld, in plaats van gedurende het verloop van een onderzoeksproject. Onderzoekers hebben baat bij een *de-novo*-aanpak, waarbij data automatisch en in realtime FAIR worden gemaakt wanneer deze verzameld worden.

In Hoofdstuk 6 presenteren we een workflow voor *de-novo* FAIRificatie: een workflow waarbij FAIRificatie-stappen opgenomen zijn in het proces van het opzetten en verzamelen van data voor een registratie of onderzoeksproject. Dit maakt geautomatiseerde, realtime FAIRificatie mogelijk, zonder tussenkomst van data management of data entry personeel. De workflow is toegepast op een registratie voor vasculaire anomalieën en kan voor een groot deel worden hergebruikt door andere onderzoeksprojecten.

In Hoofdstuk 7 beschrijven we de implementatie en evaluatie van de workflow van Hoofdstuk 6 in een Electronic Data Capture (EDC) systeem, de plaats waar medische on-

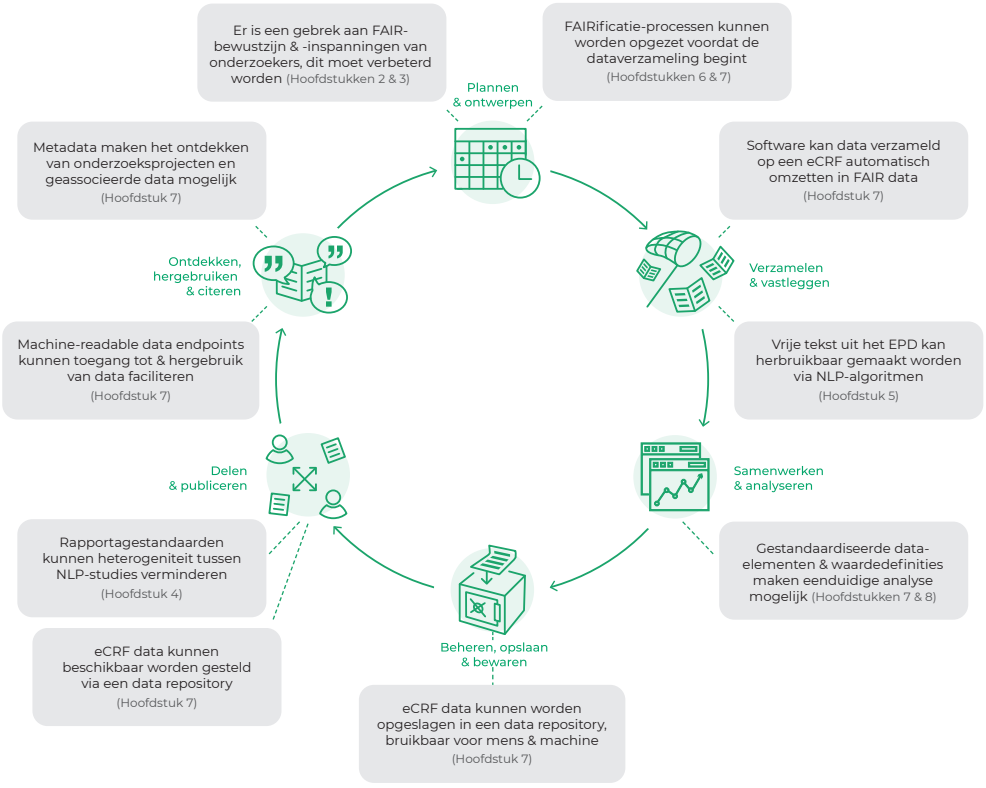
derzoeksdata vaak worden verzameld en opgeslagen via electronic Case Report Forms (eCRF's). Onze implementatie zorgt ervoor dat eCRF-data die in een EDC-systeem zijn ingevoerd, kunnen worden omgezet in machineleesbare, FAIR-data met behulp van een semantisch datamodel (een canonieke representatie van de data, gebaseerd op ontologieconcepten en semantisch-webstandaarden) en mappings van het model naar vragen op het eCRF. De toepassing ervan in de registratie voor vasculaire anomalieën en de evaluatieresultaten laten zien dat de methode kan worden gebruikt om klinische onderzoeksdata FAIR te maken wanneer ze worden ingevoerd in een eCRF, zonder tussenkomst van data management en data entry personeel.

In Hoofdstuk 8 beschrijven we het proces van het omzetten van de International Society for the Study of Vascular Anomalies (ISSVA) classificatie, een classificatie voor vasculaire anomalieën die specialisten in staat stelt diagnoses eenduidig te classificeren, in een ontologie. De classificatie was alleen beschikbaar in een formaat dat door mensen te lezen was, maar is nu machineleesbaar en interoperabel gemaakt en bevat nu verwijzingen naar bestaande ontologieën. Wij verwachten dat de ontologie door het gebruik van persistent identifiers en eenduidige definities de interoperabiliteit tussen klinische studies en registraties zal vergroten en daarom zal bijdragen aan "FAIRer" onderzoek naar vasculaire anomalieën.

Conclusie

Hoofdstuk 9 bevat de algemene discussie van het werk beschreven in dit proefschrift. De conclusies die in dit proefschrift worden getrokken, zijn samengevat in Figuur 2.

De FAIR Principles krijgen steeds meer tractie. Veel onderzoekers zijn momenteel echter niet op de hoogte van de (betekenis van de) FAIR Principles of hoe ze deze in hun werk kunnen toepassen. Het werk beschreven in dit proefschrift kwantificeert het gebrek aan FAIR-bewustzijn van klinisch onderzoekers en onderzoeksondersteunend personeel en biedt een FAIRificatie-workflow met tooling die is opgenomen in de huidige manier van werken van onderzoekers, de Research Data Life Cycle. Workflows en tools dragen bij aan het schaalbaar maken van data FAIRificatie. Gebruiksvriendelijke tools die zijn geïntegreerd in de software die door onderzoekers wordt gebruikt zijn hierin cruciaal, omdat ze geautomatiseerde FAIRificatie van data gedurende het verloop van een onderzoeksproject vergemakkelijken en het grootste deel van het werk van onderzoekers uit handen nemen. Om verdere schaalbaarheid mogelijk te maken moeten beleidsmakers, subsidieverstrekkers en onderzoeksinstituten samenwerken om standaarden, methoden, ondersteuning, tools en financiering te bieden aan de onderzoeksgemeenschap, waardoor FAIRificatie van onderzoeksdata een gezamenlijke missie wordt. Uiteindelijk baant dit de weg voor een toekomst waarin verzamelde data zoveel mogelijk waarde toevoegen aan de patiëntenzorg.



Figuur 2: Samenvatting van de belangrijkste bevindingen en conclusies, gelinkt aan de Research Data Life Cycle

FAIR: Findable, Accessible, Interoperable en Reusable, NLP: Natural Language Processing, eCRF: electronic Case Report Form, EPD: Electronisch Patiënten Dossier

Curriculum vitae

Martijn Gerard Kersloot was born in Amsterdam, the Netherlands, in 1996 at the VU University Medical Center (VUmc, now part of Amsterdam UMC). He grew up with his parents in Rijsenhout. After completing his pre-university education at the Kaj Munk College in Hoofddorp (2014), he started studying Medical Information Sciences at the University of Amsterdam and moved to Amsterdam.

During the first year of his bachelor's studies (2015) he started working at Castor, a health-tech startup that accelerates the medical research process with user-friendly technology. In the last year of his bachelor's studies, Martijn visited the University of Victoria in Victoria, BC, Canada as a co-op student and wrote his thesis under the supervision of dr. Francis Lau. After obtaining his bachelor's degree in 2017, he started a pre-PhD trajectory, a combination of the Medical Informatics Master's program and a PhD program, at the Department of Medical Informatics at the Academic Medical Center (AMC, now part of Amsterdam UMC) in collaboration with Castor. Between 2017 and 2021 he performed the research described in this thesis under the supervision of prof. dr. Ameen Abu-Hanna, dr. ir. Ronald Cornet, and dr. Derk Arts. After graduating cum laude in 2019, Martijn started working full-time on his PhD. As part of his PhD research, he spent three months abroad at The Stanford Center for Biomedical Informatics Research (BMIR), Stanford, CA, United States, and performed research under the supervision of prof. dr. Mark Musen. Next to his PhD, Martijn was a board member of Jong Amsterdam UMC – a platform connecting over 3000 young Amsterdam UMC colleagues through social and educational events.

Martijn is currently living in Aalsmeer and combines 'the best of both worlds' (business and academia) as a Product Owner at Castor and postdoctoral researcher at the Department of Medical Informatics in Amsterdam UMC.

Portfolio

PhD candidate:	Martijn G. Kersloot
Period:	September 2017 to December 2021
Supervisors:	Prof. dr. Ameen Abu-Hanna
	Dr. ir. Ronald Cornet
Co-supervisor:	Dr. Derk L. Arts

PhD training	Year	ECTS
General courses		
The AMC World of Science – AMC, Amsterdam	2017	0.7
Project Management – AMC, Amsterdam	2019	0.6
Specific courses		
Introduction to Systematic Review and Meta-Analysis – Johns Hopkins University, Coursera	2017	1.3
SNOMED CT Implementation Course – IHTSDO (online)	2017 - 2018	3.6
Clinical Epidemiology: Systematic Reviews – AMC, Amsterdam	2018	0.7
Clinical Epidemiology: Evaluation of Medical Tests – AMC, Amsterdam	2018	0.9
Clinical Epidemiology: Observational Epidemiology – AMC, Amsterdam	2018	0.6
Clinical Epidemiology: Randomized Clinical Trials – AMC, Amsterdam	2019	0.6
Modeling Biomedical Systems: Ontology, Terminology, Problem Solving – Stanford University, US	2020	0.6
ODM Implementation – CDISC (online)	2021	1.4
CDASH Implementation – CDISC (online)	2021	1.4
Seminars, workshops and master classes		
GO-FAIR Metadata4Machines Workshop – Leiden	2018	0.5
GO-FAIR FAIR Funders meeting – Leiden	2019	0.3
BMI Tuesday Talks – Stanford, US	2020	0.5
BMIR Research Colloquium – Stanford, US	2020	0.3
ELIXIR Interoperability Platform Meeting – online	2020	0.6
EJP CDE Semantic Model Hackathon – online	2020	0.4
How to work and write from home for early career researchers – online	2020	0.1

PhD training (continued)	Year	ECTS
Presentations		
Poster: "Real-time FAIRification of rare disease patient registry data" – DTL Communities @ Work conference (Utrecht)	2018	0.5
Software demo: "Real-time FAIRification of rare disease patient registry data" – SWAT4(HC)LS (Antwerp, Belgium)	2018	0.5
Oral: "The future of data standardization in medical research: FHIR & FAIR" – Webinar (online)	2019	0.5
Poster: "Real-time FAIRification of rare disease patient registry data" – HIMSS19 AMDIS/HIMSS Physicians Executive Symposium (Orlando, US)	2019	0.5
Oral: "FAIRification at the source: Transforming 'raw' eCRF data into machine-readable data" – EJP-RD CDE Semantic Model Hackathon (online)	2020	0.5
Oral: "A FAIRer Registry: Example of the Registry of Vascular Anomalies" – EJP RD General Assembly (online)	2020	0.5
Oral: "Applying the FAIR Data principles to a Rare Disease registry: a case study of the VASCA registry" – BioSB (online)	2020	0.5
Oral: "Castor's aanpak voor het FAIR maken van onderzoeksdata 'aan de bron'" – Webinar (online)	2021	0.5
Oral: "FAIRification efforts of clinical researchers: the current state of affairs" – EFMI STC (online)	2021	0.5
Oral: Promovendidag – Breukelen; Amsterdam; online	2018 - 2021	1.5
Conferences		
Medisch Informatica Congres (MIC) – Veldhoven	2017	0.5
Open Science Conference – Berlin, Germany	2018	0.5
DTL Communities @ Work – Utrecht	2018	0.25
SWAT4(HC)LS – Antwerp, Belgium	2018	0.75
HIMSS19 & Physicians' Executive Symposium – Orlando, US	2019	1.0
HL7 FHIR Dev Days – Amsterdam	2017-2019	2.25
Bioinformatics & Systems Biology conference (BioSB) 2020 – Online	2020	0.5
EFMI - Special Topic Conference (STC) – Online	2021	0.75

PhD training (continued)	Year	ECTS
Other		
Organisation Promovendidag: "KIK's Choice Awards"	2020 - 2021	1.0
Board member Jong Amsterdam UMC	2020 - 2021	1.0
Developing and updating Amsterdam Public Health Quality Handbook	2020 - 2021	0.5

Teaching	Year	ECTS
Tutoring, mentoring		
Project supervision: Uitwisseling van gezondheidsgegevens en -kennis – Bachelor Medische Informatiekunde	2021	0.5
Supervising		
1-month master Medical Informatics internship – Inez: "The use of HL7 FHIR to semantically model clinical research data"	2019	0.5
2-month master Medical Informatics internship – Sander: "Determining the variation in the modeling of electronic Case Report Form variables"	2020	1.0
5-month bachelor Medische Informatiekunde internship – Fauve: "Adoptie en kwaliteit van Common Data Elements in Nederland"	2021	2.0
Other		
Workshop: "Preparing data for machines: make data linkable & host FAIR data" – International Summer School on Rare Disease Registries and FAIRification of Data (online)	2020	0.5
Course: "Finding and capturing data" – Helis FAIR data stewardship course (online)	2021	0.5
Workshop: "Making Data Machine-readable & Hosting FAIR data: an example of a registry becoming FAIR" – International Summer School on Rare Disease Registries and FAIRification of Data (online)	2021	0.5
Workshop: "FAIR Game" – International Summer School on Rare Disease Registries and FAIRification of Data (online)	2021	0.5

List of publications

Publications in this thesis

1 FAIRification efforts of clinical researchers: the current state of affairs 2021
Kersloot MG, van Damme P, Abu-Hanna A, Cornet R, Arts DL
Studies in Health Technology and Informatics 287, 35-39. doi: 10.3233/SHTI210807

2 Data FAIRification in clinical research: perceptions and behavior of researchers and research support staff 2021
Kersloot MG, Abu-Hanna A, Cornet R, Arts DL
Submitted

3 The International Society for the Study of Vascular Anomalies (ISSVA) Ontology 2021
van Damme P *, **Kersloot MG***, dos Santos Vieira B, Schultze Kool L, Cornet R
Submitted

4 De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data 2021
Kersloot MG, Annika Jacobsen A, Groenen KHJ, dos Santos Vieira B, Kaliyaperumal R, Abu-Hanna A, Cornet R, 't Hoen PAC, Roos M, Schultze Kool L, Arts DL
Journal of Biomedical Semantics 122, 103897 (2021). doi: 10.1016/j.jbi.2021.103897

5 The de novo FAIRification process of a registry for vascular anomalies 2021
Groenen KHJ *, Jacobsen A *, **Kersloot MG**, dos Santos Vieira B, van Enckevort E, Kaliyaperumal R, Arts DL, 't Hoen PAC, Cornet R, Roos M, Schultze Kool L
Orphanet Journal of Rare Diseases 16, 376 (2021). doi: 10.1186/s13023-021-02004-y

6 Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies 2020
Kersloot MG, Van Putten FJP, Abu-Hanna A, Cornet R, Arts DL
Journal of Biomedical Semantics 11, 14 (2020). doi: 10.1186/s13326-020-00231-z

7 Automated SNOMED CT concept and attribute relationship detection through a web-based implementation of cTAKES 2019
Kersloot MG, Lau F, Abu-Hanna A, Arts DL, Cornet R
Journal of Biomedical Semantics 10, 14 (2019). doi: 10.1186/s13326-019-0207-3

** Equal contributors.*

Other publications

8 Applying the FAIR Data Principles to the Registry of Vascular Anomalies (VASCA) 2020
Dos Santos Vieira B, Groenen K, 't Hoen PAC, Jacobsen A, Roos M, Kaliyaperumal R, **Kersloot MG**, Cornet R, Schultze Kool L
Studies in Health Technology and Informatics 271, 115-116 (2020). doi: 10.3233/SHTI200085

9 The Need of Industry to Go FAIR 2020
van Vlijmen H, Mons A, Waalkens A, Franke W, Baak A, Ruiter G, Kirkpatrick C, Bonino da Silva Santos LO, Meerman V, Jellema R, Arts DL, **Kersloot MG**, Knijnenburg SL, Lusher S, Verbeeck R, Neefs J
Data Intelligence 2 (1-2), 276-284 (2020). doi: 10.1162/dint_a_00050

10 FAIR Principles: Interpretations and Implementation Considerations 2020
Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo CT, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa K, Hooft RWW, Imming M, Jeffery KG, Kaliyaperumal R, **Kersloot MG**, Kirkpatrick CR, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S, Bonino da Silva Santos LO, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson MD, Willighagen EL, Wittenburg P, Roos M, Mons B, Schultes E
Data Intelligence 2 (1-2), 10-29 (2020). doi: 10.1162/dint_r_00024

List of contributing authors

Ameen Abu-Hanna
Department of Medical Informatics
Amsterdam Public Health Research Institute
Amsterdam UMC, University of Amsterdam
Amsterdam, The Netherlands

Derk L. Arts
Castor
Amsterdam, The Netherlands

Ronald Cornet
Department of Medical Informatics
Amsterdam Public Health Research Institute
Amsterdam UMC, University of Amsterdam
Amsterdam, The Netherlands

Philip van Damme
Department of Medical Informatics
Amsterdam Public Health Research Institute
Amsterdam UMC, University of Amsterdam
Amsterdam, The Netherlands

Esther van Enckevort
Department of Genetics &
Genomic Coordination Center
University of Groningen,
University Medical Center Groningen
Groningen, The Netherlands

Karlijn H.J. Groenen
Department of Medical Imaging
Radboud Institute for Health Sciences
Radboud university medical center
Nijmegen, The Netherlands

Peter A.C. 't Hoen
Center for Molecular & Biomolecular Informatics
Radboud Institute for Molecular Life Sciences
Radboud university medical center
Nijmegen, The Netherlands

Annika Jacobsen
Department of Human Genetics
Leiden University Medical Center
Leiden, The Netherlands

Rajaram Kaliyaperumal
Department of Human Genetics
Leiden University Medical Center
Leiden, The Netherlands

Francis Lau
School of Health Information Science
University of Victoria
Victoria, Canada

Florentien J.P. van Putten
Department of Medical Informatics
Amsterdam Public Health Research Institute
Amsterdam UMC, University of Amsterdam
Amsterdam, The Netherlands

Marco Roos
Department of Human Genetics
Leiden University Medical Center
Leiden, The Netherlands

Bruna dos Santos Vieira
Department of Medical Imaging
Radboud Institute for Health Sciences
Radboud university medical center
Nijmegen, The Netherlands

Center for Molecular & Biomolecular Informatics
Radboud Institute for Molecular Life Sciences
Radboud university medical center
Nijmegen, The Netherlands

Leo Schultze Kool
Department of Medical Imaging
Radboud Institute for Health Sciences
Radboud university medical center
Nijmegen, The Netherlands

Authors' contributions

Chapter 2

Martijn G. Kersloot
Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft

Ronald Cornet
Conceptualization, Methodology, Writing - Review & Editing

Derk L. Arts
Conceptualization, Methodology, Writing - Review & Editing

Ameen Abu-Hanna
Conceptualization, Methodology, Writing - Review & Editing

Chapter 3

Martijn G. Kersloot
Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft

Philip van Damme
Methodology, Data Curation, Writing - Review & Editing

Ronald Cornet
Conceptualization, Methodology, Writing - Review & Editing

Derk L. Arts
Conceptualization, Methodology, Writing - Review & Editing

Ameen Abu-Hanna
Conceptualization, Methodology, Writing - Review & Editing

Chapter 4

Martijn G. Kersloot
Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft

Florentien J.P. van Putten
Formal analysis, Investigation, Data Curation

Ronald Cornet
Conceptualization, Methodology, Writing - Review & Editing

Derk L. Arts
Conceptualization, Methodology, Writing - Review & Editing

Ameen Abu-Hanna
Conceptualization, Methodology, Writing - Review & Editing

Chapter 5

Martijn G. Kersloot
Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft

Francis Lau
Conceptualization, Methodology, Writing - Review & Editing

Ronald Cornet
Conceptualization, Methodology, Writing - Review & Editing

Derk L. Arts
Methodology, Writing - Review & Editing

Ameen Abu-Hanna
Methodology, Writing - Review & Editing

Chapter 6

Karlijn H.J. Groenen
Conceptualization, Methodology, Writing - Original Draft

Annika Jacobsen
Conceptualization, Methodology, Writing - Original Draft

Martijn G. Kersloot
Conceptualization, Methodology, Software, Writing - Original Draft

Bruna dos Santos Vieira
Conceptualization, Methodology, Writing - Original Draft

Esther van Enckevort
Conceptualization, Methodology, Writing - Review & Editing

Rajaram Kaliyaperumal
Conceptualization, Methodology, Writing - Review & Editing

Derk L. Arts
Conceptualization, Methodology, Writing - Review & Editing

Peter A.C. 't Hoen
Conceptualization, Methodology, Writing - Review & Editing

Ronald Cornet
Conceptualization, Methodology, Writing - Review & Editing

Marco Roos
Conceptualization, Methodology, Writing - Review & Editing

Leo Schultze Kool
Conceptualization, Methodology, Writing - Review & Editing

Chapter 7

Martijn G. Kersloot
Conceptualization, Methodology, Software, Writing - Original Draft

Annika Jacobsen
Conceptualization, Methodology, Writing - Review & Editing

Karlijn H.J. Groenen
Conceptualization, Methodology, Writing - Review & Editing

Bruna dos Santos Vieira
Writing - Review & Editing

Rajaram Kaliyaperumal
Writing - Review & Editing

Ameen Abu-Hanna
Supervision, Writing - Review & Editing

Ronald Cornet
Conceptualization, Supervision, Writing - Review & Editing

Peter A.C. 't Hoen
Writing - Review & Editing

Marco Roos
Conceptualization, Supervision, Writing - Review & Editing

Leo Schultze Kool
Conceptualization, Supervision, Writing - Review & Editing

Derk L. Arts
Conceptualization, Supervision, Writing - Review & Editing

Chapter 8

Philip van Damme
Conceptualization, Methodology, Software, Writing - Original Draft

Martijn G. Kersloot
Conceptualization, Methodology, Software, Writing - Original Draft

Bruna dos Santos Vieira
Conceptualization, Methodology, Software, Writing - Review & Editing

Leo Schultze Kool
Conceptualization, Writing - Review & Editing

Ronald Cornet
Methodology, Writing - Review & Editing

The FAIR Principles, stating that research data and metadata should be Findable, Accessible, Interoperable, and Reusable for both humans and machines, are experiencing a vast uptake in acceptance and implementation by researchers, research institutes, funders, and government bodies. However, many researchers are currently unaware of the FAIR Principles, their implications, or how they can be applied to their research. Furthermore, existing workflows to make data FAIR are designed to be executed after research projects have been conducted and data have been collected, rather than throughout the life cycle of research projects.

The work presented in this thesis provides insight into researchers' and research support staff's knowledge and perspectives on the implementation of the FAIR Principles in practice (Part I), determines the role of Natural Language Processing in making data more FAIR (Part II), and develops a process for making data FAIR from the beginning of a research project and at the source (Part III). The presented work contributes to a future in which FAIR research data are the default, and the process of making data FAIR is optimized, to add maximum value to patient care with minimal cost, effort, and delay.