Defining the quality of data within rare disease registries: A systematic review

Jessica Anderson, Malika Alimussina, S. Faisal Ahmed
Office for Rare Conditions, School of Medicine, Dentistry & Nursing, University of Glasgow, UK

doi: 10.56012/knxk2392

Correspondence to:
S. Faisal Ahmed
Faisal.ahmed@glasgow.ac.uk

Abstract

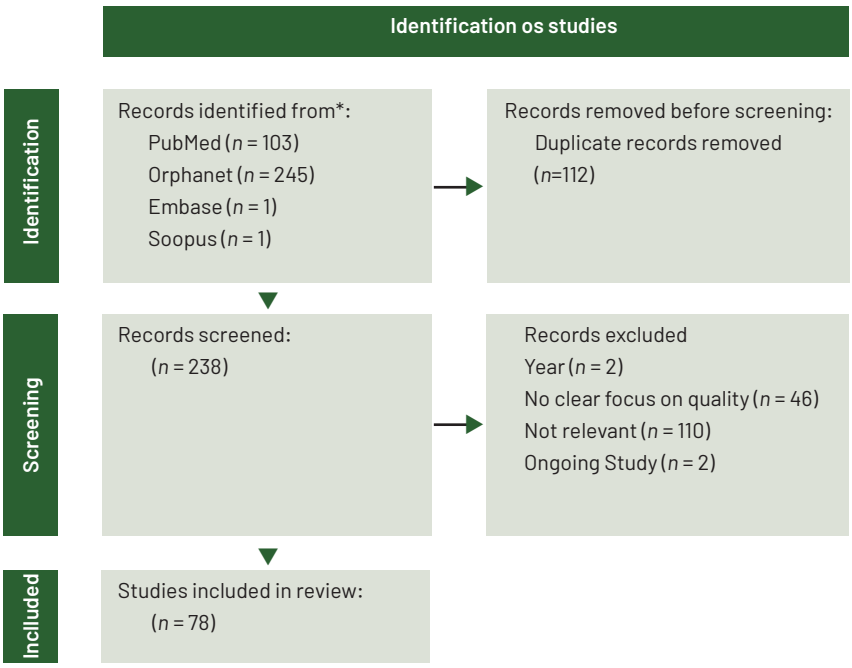
Rare diseases have a low prevalence within society, resulting in limited awareness and challenges with data availability for research. While rare disease registries offer valuable data, ensuring quality of the data is essential. This review explores key themes and influencing factors affecting data quality in rare disease registries. Studies were identified through a pre-defined search term across multiple databases and screened for recurring themes and terms. The findings indicate a growing emphasis on data quality and evolving perspectives on how it is defined and assessed through the years.

Rare diseases are often defined as conditions that affect fewer than 5 out of 10,000 members of the general population[1] and these conditions may affect up to 6%-7% of the world's population.[2] The low prevalence of these conditions often leads to limited awareness of the conditions as well as their management among both the public and healthcare professionals.[2] The lack of data affects the development of an adequate amount of evidence that can inform safety and effectiveness of drugs, diagnosis, and research in general. To address these challenges, one possible solution is the development of patient registries.

Patient registries are databases that are designed to systematically collect, store, and analyse clinical data. They can be used to track patient demographics, diagnosis, treatments, and outcomes, enabling longitudinal studies to take place on a large scale. The data that are collected in these registries often represent a setting that is beyond what is encountered in controlled clinical trials or experimental environments. So these patient registries not only address the challenges of limited and heterogenous data, but they also collect real-world data that reflects how people utilise healthcare services and respond to interventions in their everyday lives.[3,4] Real-world data provide insights into disease progression, treatment outcomes, and patient experiences, which are essential for informing healthcare policy, improving clinical care, development of new drugs and interventions, monitoring the use of these interventions and for performing comparative effectiveness research.[5] As rare disease registries start to play an increasingly pivotal role in rare disease research, the rare disease community has seen a proliferation of these registries and this has an implication on long-term sustainability of these platforms.

The critical factor that will influence the long-term sustainability of a rare disease registry will be its quality and this can be broadly divided into two categories. The first one relates to its operation systems and the second category, which is equally important, relates to the data that the registry collects.[6] This is even more important in rare diseases where the populations are very small and poor data quality may skew the results or lead to inconclusive results thus limiting the acceptability of the findings. Data quality itself may be defined in several ways including completeness, interoperability, accuracy, validity, consistency, timeliness, uniqueness and traceability.[7,8] Amongst existing registries, it is clear that the definition of registry quality may be quite variable[9] and the level of consensus that may exist for data quality is also unclear. It is important to understand the key concepts of data quality so that resources can be directed towards these to ensure long-term sustainability. Furthermore, registries with a higher level of data quality are more likely to have greater acceptability amongst health care providers. The current systematic review was, therefore, performed to explore the key concepts of data quality that are reported in contemporary rare disease registry literature.

Figure 1. PRISMA flowchart outlining the inclusion exclusion criteria used to screen the literature



Methods

A systematic review was performed to examine how data quality is defined in rare disease registries by synthesising literature from 2010 to 2025 and identifying key themes and related components that define data quality. Thematic analysis was performed to categorise recurring themes and trends that were observed within the literature. The inclusion criteria included publications that were published in English in a peer reviewed journal from 2010 onwards and had a clear focus on data quality and rare disease registries. The 15-year time period was chosen as it was felt to be a relevant period to capture a sufficient amount of literature within the field. Rare diseases were included in the criteria to ensure the relevant population was captured appropriately. Non-peer reviewed literature was excluded to ensure the reliability of the literature for this analysis. The systematic review was conducted and reported in accordance with the method outlined in the Cochrane Handbook for Systematic Reviews[10] and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[11] Literature search results were uploaded to Covidence (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia). Articles were manually screened by title and abstract to determine eligibility according to the inclusion criteria above. Relevant full-text studies were collated and evaluated for eligibility for inclusion (Figure 1). The selected studies were also screened for the definitions and themes as previously described.[12] These data were then extracted from Covidence for frequency analysis of the definitions of data quality and the factors that affect data quality. Lastly, thematic analysis was performed to identify recurring themes within the literature. Following initial familiarisation with the literature within the field, key concepts and phrases were identified (Figure 2). The thematic analysis was used to identify trends in defining data quality in rare disease registries and trends in factors that may influence data quality in rare disease registries over the last 15 years. These temporal trends were arbitrarily divided into four time periods of three years each. The co-occurrence of themes was analysed using R, employing the tidyverse, igraph, and ggraph packages. Each article was assigned a unique article ID to facilitate tracking. Themes associated with each article ID were identified, and pairwise co-occurrences of themes within individual articles were computed. These co-occurrences were then aggregated across all articles to assess the frequency of theme co-occurrence throughout the data set.

Table 1. Themes and terms

Theme	Term
Completeness	Complete Completed Completeness
Selection Bias	Bias Selection bias
Validity	Validity Valid Validate
Accuracy	Accuracy Accurate
Interoperability Duplication	Interoperability Duplicate Duplication
Standardisation	Standardisation Standardised
Common Data Set Elements	Common data set elements MDS

Themes and their corresponding codes. It is important to note that for standardisation both the American and British spellings were used to screen the literature.
Abbreviation, MDS, minimum data set

Results

Frequency of definitions of data quality

A total of 78 studies were included, and within these studies 9 themes were identified: completeness, selection bias, validity, accuracy, consistency, interoperability, duplication, standardisation, and common data set elements. These 9 themes were further subdivided into terms that represented those that were used within these themes (Table 1). On the other hand, terms such as common data set elements, minimum data set (MDS) were not very frequent.

Trends in definitions of data quality

The total number of term occurrences grew steadily from 50 in 2010–2013 to 876 in 2022–2025, representing a 17.5-fold increase over the study period (Table 2). Terms related to completeness (e.g. *completeness*, *complete*, *completed*) were among the most frequently cited, with *completeness* alone appearing 224 times, followed by *complete* (166 times) and *completed* (90

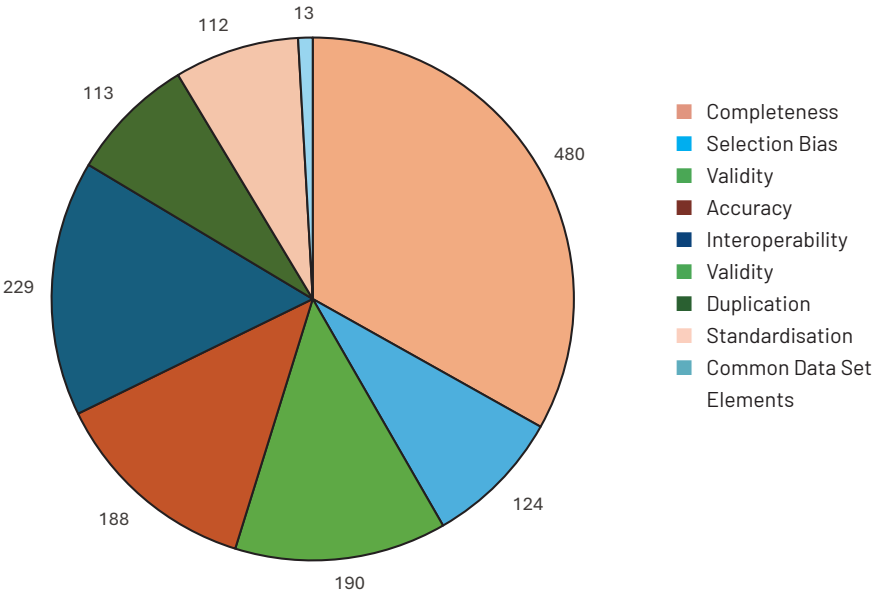


Figure 2. Theme frequency pie chart describing the frequency of terms used to screen the literature

Table 2. Temporal trends in the reporting of themes and terms, 2010–2025

Theme	Term	Temporal Group 2010–2013	2014–2017	2018–2021	2022–2025	Total for Term
Accuracy	Accuracy	0	26	30	59	115
	Accurate	0	13	24	36	73
	Total	0	39	54	95	188
Selection Bias	Bias	2	14	17	55	88
	Selection Bias	8	6	5	17	36
	Total	10	20	22	72	124
Common Data Set Elements	Common data set elements	0	0	0	0	0
	MDS	0	0	6	7	13
	Total	0	0	6	7	13
Completeness	Complete	13	20	37	96	166
	Completed	15	7	20	48	90
	Completeness	1	66	40	117	224
	Total	29	93	97	261	480
Consistency	Consistency	2	11	26	58	97
	Consistent	2	6	17	47	72
	Total	4	17	43	105	169
Duplication	Duplicate	0	21	44	14	79
	Duplication	1	7	13	13	34
	Total	1	28	57	27	113
Interoperability	Interoperability	0	5	79	145	229
Standardisation	Standardisation	1	2	12	15	30
	Standardised	2	9	17	54	82
	Total	3	11	29	69	112
Validity	Valid	0	4	12	12	28
	Validate	2	3	15	17	37
	Validity	1	38	20	66	125
	Total	3	45	47	95	190
Total Terms		**50**	**258**	**434**	**876**	**1618**

Abbreviation: MDS, minimum data set

times). Similarly, *interoperability* experienced a marked increase, from no mentions before 2014 to 145 mentions in 2022–2025, making it the most cited individual term overall (229 total mentions). Conversely, certain terms such as *common data set elements* and *MDS* were rarely mentioned or not at all, indicating either limited focus or a preference for alternative terminology. With the exception of the term *common data set elements* phrase, all other terms were present from 2018 onwards (Figure 3). From 2010 to 2013, the most frequently occurring terms were *completed* and *completeness*, marking completeness as the dominant theme in that early period. In the following period, 2014–2017, *completeness* remained the most frequent term, but it was followed closely by *validity*. A more marked shift occurred in 2018–2021, when *interoperability* and *duplicate* became the most frequently mentioned terms. By 2022–2025, the top two terms were again *interoperability* and *completeness*.

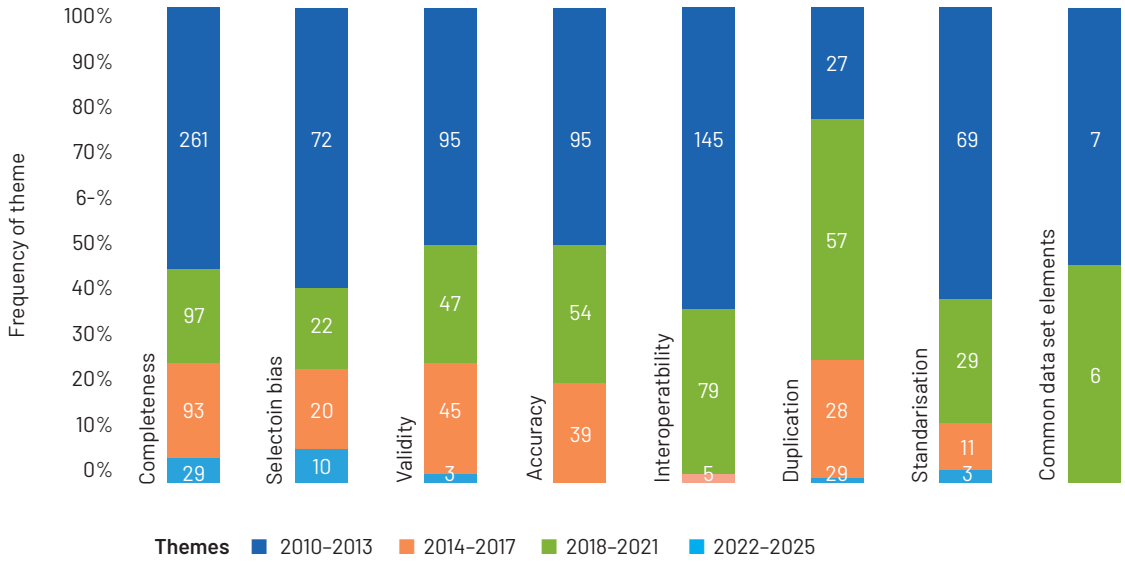


Figure 3. The frequency of themes for each of the year groups previously defined

Relationship of themes to each other

All themes co-occurred with at least one other, demonstrating that each theme had been discussed alongside others at some point in the literature (Figure 4). Whilst the theme *common data set elements* did not have a high overall frequency, it was still well interconnected. This was because the articles that had this theme also had multiple other themes occurring at the same time as well. This means that whilst the theme overall was not frequent in the literature, it was interconnected with the other themes.

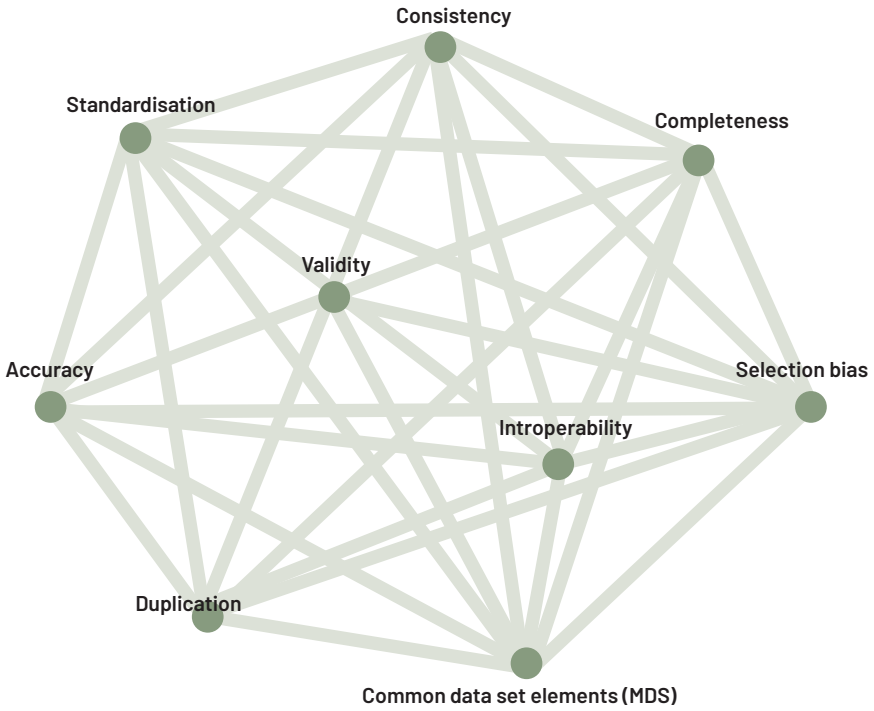


Figure 4. Theme network graph showing how interlinked each of the themes are with one another.

Discussion

This review set out to explore how data quality is defined within rare disease registries by analysing literature published between 2010 and 2025, with the aim of identifying key themes and influencing factors. Using thematic analysis framework,[12] nine recurring themes were identified across the included studies: completeness, selection bias, validity, accuracy, consistency, interoperability, duplication, standardisation, and common data set elements. Together, these themes reflect the complexity of data quality and the range of priorities currently shaping the field.

The findings show a clear progression in how data quality has been approached over time. Between 2010 and 2013, the focus tended to be on more basic aspects of quality – particularly completeness and whether data had been fully recorded – highlighting an early concern with ensuring registries captured the full picture. From 2018 onwards, however, the emphasis has shifted towards more system-level issues such as interoperability and duplication. This change points to a deeper and more technical understanding of what makes data useful, particularly when it is shared across settings or used for secondary purposes. The growing frequency of terms over time reflects an increasing interest in defining and improving data quality across both academic and clinical contexts. In addition, the overlap between themes – demonstrated through co-occurrence – suggests that these concepts are not being

Image: Freepik

considered in isolation, but as part of a broader, interrelated understanding of quality. This highlights the interconnectedness of these concepts and suggests that the definition of data quality within rare disease registries is inherently multidimensional.

Health care professionals face a variety of barriers to participating in rare disease registries. Many health care professionals are not aware of rare disease registries and even when they are aware of these registries their level of participation is limited.[13] Clinicians and associated administrative and care staff often have heavy workloads, leaving little time for data entry or patient follow-up.[14] If data were sufficiently interoperable, they could flow between different sources and the need for manual entry that may also lead to transcription errors could be minimised. However, even if this was possible, it is likely that at an institutional level, without local approval, free data flow for highly sensitive data will be challenging. Rare disease registries rarely need to collect that are real-time, and a solution for addressing the time constraints is to develop systems that can bulk download source data and subsequently upload the data at a time that is convenient. However, this still requires the need to agree on standardised data sets that can be collected universally. These data sets are referred to in different ways in the literature including common data elements,[15] core outcome sets,[16] and minimum data sets.[17] By minimising the amount of data that is collected in rare disease registries, projects such as GloBE-Reg, a global registry for novel therapies in rare bone and endocrine conditions, are aiming to improve the data quality.[17] One potential limitation of this study is the possibility that not all relevant terms such as *common data elements* or *minimum data set* frequencies were captured. This is likely due to the terms being used to search the literature not capturing the frequency of these themes accurately, potentially introducing bias.

Overall, the findings from this review highlight both an increased focus on data quality in rare disease registries over time and a shift in how quality is being conceptualised. While earlier studies primarily emphasised completeness and validity, more recent literature places greater attention on themes such as interoperability, duplication, and consistency. This shift suggests a growing and more nuanced understanding of what makes data high quality.

Disclosures and conflicts of interest

The authors are supported by GenSci, Novo Nordisk and Pfizer for the GloBE-Reg project. The authors declare no conflicts of interest.

Data availability statement

For inquiries about data and other supplemental information, please contact the corresponding author.

Refernces

1. European Commissions. Rare diseases - European Commission. 2023 [cited 2025 Jan 29]. Available from: https://health.ec.europa.eu/rare-diseases-and-european-reference-networks/rare-diseases_en
2. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165–73. doi:10.1038/s41431-019-0508-0
3. Hageman IC, van Rooij IALM, de Blaauw I, et al. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. Orphanet J Rare Dis. 2023;18(1):106. doi:10.1186/s13023-023-02719-0
4. Shourick J, Wack M, Jannot A. Assessing rare diseases prevalence using literature quantification. Orphanet J Rare Dis. 2021;16(1):139. doi:10.1186/s13023-020-01639-7
5. Raycheva R, Kostadinov K, Mitova E, et al. Challenges in mapping European rare disease databases, relevant for ML-based screening technologies in terms of organizational, FAIR and legal principles: scoping review. Front Public Health. 2023;11: 1214766. doi:10.3389/fpubh.2023.1214766
6. Kodra Y, Weinbach J, Posada-De-La-Paz M, et al. Recommendations for improving the quality of rare disease registries. Int J Environ Res Public Health. 2018;15(8):1644. doi:10.3390/ijerph15081644
7. Wang J, Liu Y, Li P, et al. Overview of data quality : examining the dimensions, antecedents, and impacts of data quality. J Knowl Econ. 2024;15:(1): 1159–78. doi:10.1007/s13132-022-01096-6

8. Jonker CJ, de Vries ST, van den Berg HM, et al. Capturing data in rare disease registries to support regulatory decision making: a survey study among industry and other stakeholders. Drug Saf. 2021;44(8):853–61. doi:10.1007/s40264-021-01081-z
9. Ali SR, Bryce J, Kodra Y, Taruscio D, et al. The quality evaluation of rare disease registries-an assessment of the essential features of a disease registry. Int J Environ Res Public Health. 2021;18(22):11968. doi:10.3390/ijerph182211968
10. Higgins JPT, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). doi:10.1002/9781119536604
11. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
12. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach. 2020;42(8):846–54. doi:10.1080/0142159x.2020.1755030
13. Ali SR, Bryce J, Cools M, et al. The current landscape of European registries for rare endocrine conditions. Eur J Endocrinol. 2019;180(1):89–98. doi:10.1530/eje-18-0861
14. Kyriakou A, Dessens A, Bryce J, et al. Current models of care for disorders of sex development - results from an international survey of specialist centres. Orphanet J Rare Dis. 2016;11:155. doi:10.1186/s13023-016-0534-8
15. Ali SR, Bryce J, Smythe C, et al. Supporting international networks through platforms for standardised data collection-the European Registries for Rare Endocrine Conditions (EuRRECa) model. Endocrine. 2021;71(3):555–60. doi:10.1007/s12020-021-02617-0
16. Flück C, Nordenström A, Ahmed SF, et al. Standardised data collection for clinical follow-up and assessment of outcomes in differences of sex development (DSD): recommendations from the COST action DSDnet. Eur J Endocrinol. 2019;181(5):545–64. doi:10.1530/eje-19-0363
17. Chen SC, Bryce J, Chen M, et al. Development of a minimum dataset for the monitoring of recombinant human growth hormone therapy in children with growth hormone deficiency: a GloBE-Reg Initiative. Horm Res Paediatr. 2024; 97(4):365–73. doi:10.1159/000533763

Author information

S. Faisal Ahmed, MD, FRCPCH, received his clinical and scientific training in Edinburgh, London, and Cambridge. He was appointed as a consultant in paediatric endocrinology at the Royal Hospital for Sick Children, Yorkhill, Glasgow, in 2000 and was appointed as the Samson Gemmell Chair of Child Health at the University of Glasgow in 2012. Since 2019, he has held an additional chair at Leiden University as Professor of Endocrine Registries. His research has focussed on improving the health of people with a wide range of rare conditions, and especially those with conditions affecting sex development or growth and skeletal development. He is the Editor-in-Chief of the journal Endocrine Connections.
ORCID: 0000-0003-0689-5549

Malika Alimussina received her clinical and academic training in paediatric endocrinology in Kazakhstan and later completed her PhD in Medicine at the University of Glasgow in 2023. Dr Alimussina is currently a Clinical Scientist at the Office for Rare Conditions, University of Glasgow, where she contributes to the development and analysis of international registries for rare endocrine disorders. Her research focuses on improving long-term outcomes in individuals with Disorders of Sex Development and other rare growth and endocrine conditions. Dr Alimussina is also a faculty member at one of the European Society for Paediatric Endocrinology training schools, where she teaches and mentors early-career clinicians and researchers.

Jessica Anderson received a BSc (Hons) in Pharmacology from the University of Aberdeen in 2023 and an MSc in Precision Medicine and Pharmacological Innovation from the University of Glasgow in 2024. She is working on a PhD focused on data quality within rare disease registries, with particular emphasis on completeness, interoperability, and consistency of core clinical variables. Her research aims to develop approaches to improve the reliability of real-world data and support translational research and clinical decision-making in rare diseases.