RD ACTION
DATA AND POLICIES FOR RARE DISEASES

Recommended Practices for Data Standardisation in the Context of the operation of European Reference Networks

2017

RD-ACTION Output

Co-funded by
the Health Programme
of the European Union

Document History

This document originated as an output of the Workshop organised by RD-ACTION and co-hosted by DG SANTE, which took place in Brussels on 25th and 26th April 2017. An initial draft was shared with the topic-specific experts who attended the workshop, with a deadline for feedback of 31st July. This led to preparation of an advanced draft, which was shared with the full group of workshop participants on 8th August. Feedback was received up until 22nd September and was used to refine and generate this version of the document.

European Reference Networks and the Opportunities they afford

European Reference Networks (ERNs) are networks connecting providers of highly specialised healthcare, united for the purposes of improving access to diagnosis, treatment and high-quality care for patients with conditions requiring a particular concentration of resources or expertise. Composed of healthcare providers (HCPs) able to demonstrate the highest levels of care and research excellence, there are currently 24 approved ERNs, each dedicated to a broad rare disease area/highly specialised intervention. Almost 1000 units across 370 hospitals in 26 European countries[1] are involved as direct (full) members, with access from 2018 onwards for 'affiliated' partners (to enable the participation of countries without a full member in any given network).

At the heart of the ERN concept is the principle that wherever possible (and appropriate), expertise will travel rather than the patients themselves. In practice, this will entail a significant degree of virtual healthcare provision, which demands the exchange and accessibility of data.

In view of their dual focus on both care and research, ERNs offer an unprecedented opportunity to collect data concerning two broad 'categories' of patients whose conditions require a concentration of expertise and specialists:

- patients formally referred for virtual care/shared care under an ERN; but also;
- patients attending clinics in one of the member HCPs of an individual ERN (and possibly also 'affiliated' centres), even if not referred for virtual care under the Network .

Collecting data in a standardised manner will allow it to become syntactically and semantically interoperable, which increases the power of that data in several ways. Professionals participating in virtual patient reviews will benefit from an ability to receive information in a standardised form, as this ensures that the same terms are understood in the same way by those receiving the data (for example, diseases coded according to the same nomenclature; lab reports generated in accordance with the same reporting standards; clinical terms harmonised, etc.). Virtual review of patients, whether real-time or not, is time-consuming[2] : it is necessary to find a way to make these consultations as efficient as possible, and gaining consensus as to which data the experts will review and how they can expect to receive it/how they should provide it, is logical. Arguably however, the greater benefit -given the ERN focus on rare diseases and procedures which are classed as highly specialised- is that when data is collected in a certain way, using recommended tools and standards, it can be *re-used* and thus achieves a 'life-span' beyond the initial purpose of direct care delivery. Once pseudonymised, data can be pooled to advance diagnostics, knowledge, and understanding of the disease and of its accompanying symptoms etc. Moreover, standardization allows information held inside multiple locations to be interrogated and produce aggregate results without requiring sensitive data to travel.

[1] For details of membership per ERN and per country, see https://ec.europa.eu/health/ern/policy_en

[2] RD-ACTION, Workshop Report 'Exchanging data for virtual care within the ERN Framework' , p. 12-16 (http://www.rd-action.eu/wp-content/uploads/2016/12/Report-of-RD-ACTION-Workshop-Exchanging-Data-for-Virtual-Care-within-the-ERN-Framework-1.pdf)

Background to the Generation of the Current Document

In view of these considerations, the final session of the RD-ACTION workshop 'Exchanging data for virtual care in the ERN framework' -which took place September 27-28th 2016- outlined the following conclusions:[3]

- *The ability to share and pool data, or interrogate information across resources, is essential in the RD field, and in all fields requiring a specific concentration of expertise: only through access to a congregation of data can one attain a critical mass, which generates knowledge and drives forwards improvements in healthcare*
- *Use of agreed ontologies such as the ORDO (Orphanet Rare Disease Ontology) and the HPO (Human Phenotype Ontology) adds value to data, especially in terms of the reusability of that data – standards which exist already and have gained a certain level of 'acceptance' in the wider RD and specialised healthcare field should be promoted in the ERN framework, to enhance the value of the data which is collected, exchanged, and retained.*
- *In recommending standards for use with ERN-related data, it is important to note that the process is not unidirectional: what other standards should be embraced, which are used widely in the RD and/or specialised healthcare/ technology field and have been proven to enhance the utility of information/data (e.g. standards around coding medical devices)?*
- *It would be useful to arrange a more hands-on demonstration for some of these tools, to explore what needs to be in place, how one can use Orphanet Nomenclature, HPO, Identifiers etc. to add value to the data and increase its interoperability through FAIR[4] approaches, for instance.*
- ***It would be logical to produce a list of consensus recommendations on standardising data in ERNs***

To expand upon the points above, and to deliver upon the final conclusion, **a dedicated workshop was organised by RD-ACTION on 26-27th April 2017, co-hosted with DG SANTE.**

Four key resources/approaches were highlighted - these are not exhaustive and should be viewed as a starting point towards optimising the value of data in the ERN community. Homogeneity of *approach* is most important here: ERNs are as heterogeneous as the diseases and procedures in which they specialise, and consequently a 'one-size-fits-all' approach is not feasible. However, applying lessons of the global rare disease and data stewardship communities could pay dividends here, and enhance the power of the data -and thus of the ERNs themselves- exponentially. By using mature, consensus standards and ontologies to capture diseases and phenotypes, and agreeing additional ontologies to record *other* types of clinically-relevant information, the data which is collected, exchanged and stored by an ERN can be semantically interoperable with data from all other ERNs, but also with external rare disease, public health, and life science datasets: data can be interrogated across many lines of commonalities.

The following suggested practices were discussed and are hereby presented as an initial set of data standardisation principles to optimise the utility and reusability of data in the ERN sphere.

[3] *Ibid*. p27-8

[4] Findable, Accessible, Interoperable, Reusable for humans and computers (see below, section 4) – the FAIR data concept in fact includes a focus on each of the other 3 sections highlighted in this document; for instance, FAIR data services would typically include advice and training on using ontologies and identifiers/pseudonymisation techniques.

SECTION 1: CODING RARE DISEASES WITHIN ERNs

i. When providing diagnostics, treatment and care for a patient with a rare disease, it is important to capture the **specific** condition -i.e. the clinically distinct subset- and not merely the overall disease group.[5] Only through use of an appropriate classification system is it possible to make an accurate assessment of the actual numbers of patients afflicted with a given rare condition in Europe.

ii. The OrphaCode has been approved on both the European[6] and global[7] levels as the most appropriate nomenclature for the clinical coding of rare diseases, in view of its granularity and ability to distinguish between specific rare diseases, which makes it preferable to all alternatives.[8]

iii. The OrphaCode is regularly cross-referenced and harmonised with more mainstream systems of classification, including ICD-10, ICD-11, Snomed-CT, UMLS and OMIM – therefore, using the OrphaCode ensures a *link* to the more mainstream nomenclatures, whilst enhancing the power of the data exponentially. In some cases, use of such systems **alongside** the OrphaCode may be very beneficial (for instance, use of OMIM to capture genetic /molecular entities can be advantageous where clinical diagnosis is unclear)

iv. As ERNs -and, increasingly, their constituent HCPs and 'affiliated' partners – deal with *electronic* patient data, it is necessary to utilise an ontology (i.e. a computer-readable form of the nomenclature) to code the diseases encountered. The Orphanet Rare Disease Ontology (ORDO) has been built using the OrphaCode, and enables computer systems to understand how diseases relate to one another under a hierarchical 'tree-and-branch' structure.

v. If all ERNs use the OrphaCode (most importantly via the Clinical Patient Management System[9]), then diseases are unambiguously identifiable across corresponding data from other ERNs: data which are pooled can be -crucially- understood as pertaining to the same disease.

vi. Several ERNs are dedicated to rare cancers, which have their own very important codification system – the International Classification of Childhood Cancer (3rd edition) classifies neoplasms into twelve main diagnostic groups.

vii. Increasing the visibility of rare diseases in ERNs through adoption of the Orphanet nomenclature may, in time, support the wider dissemination and use of the OrphaCode in health information systems of Europe - dedicated guidance is available [10]

RECOMMENDATIONS FOR ERNs AND THEIR CONSTITUENT CENTRES

1. ERNs -and, as far as possible, their member HCPs and 'affiliated' partners in the course of their broader, daily activities- should promote and utilise the OrphaCode as the preferred nomenclature for capturing the suspected or confirmed diagnosis of a rare disease.

[5] Which might be acceptable in different coding circumstances, for instance, a reimbursement scenario based upon the DRG or 'Diagnosis Related Group'

[6] The OrphaCode is the subject of dedicated recommendations issued by the Commission Expert Group on Rare Diseases: *Recommendation on Ways to Improve Codification for Rare Diseases in Health Information Systems* (2014) http://ec.europa.eu/health//sites/health/files/rare_diseases/docs/recommendation_coding_cegrd_en.pdf

[7] The OrphaCode has received 'IRDiRC Recommended Resources' label, reserved for resources which "if used more broadly, would accelerate the pace of discoveries and translation to clinical services": http://www.irdirc.org/activities/irdirc-recognized-resources/

[8] For instance, SNOMED-CT classifies diseases clinically, but covers fewer than half of the diseases in Orphanet. OMIM classifies rare disease genetically, yet only 57% are incorporated.

[9] Described in Tender SANTE/2016/A4/013 (http://ted.europa.eu/udl?uri=TED:NOTICE:205468-2016:TEXT:EN:HTML)

[10] Specific guidance for countries is in preparation by WP5 of RD-ACTION http://www.rd-action.eu/news/standard-procedure-and-guide-for-coding-with-orphacodes-available/

2. All electronic systems for the capture and exchange of data should embed [11] the ORDO and should provide a field in which to capture the OrphaCode, accompanied by a second field allowing the user to indicate confidence in the accuracy of that diagnosis, by selecting either 'confirmed' or 'suspected' or else "Undetermined diagnosis.
3. Each ERN should, at some stage, consider reviewing the existing Orphanet nomenclature relative to its thematic grouping (i.e. disease or procedural area) – ideally, ERNs should envisage:
    a. establishing Working Groups/ Transversal groups on coding
    b. establishing guidelines on how to code diseases under the heading/subdomain
    c. contributing to the improvement and curation of the nomenclature, particularly through the Orphanet Knowledge Management System[12]
4. To capture the specificities of **rare cancers**, and ensure interoperability with data from the broader cancer field, ERNs should also have capacity to capture the codes used by the International Classification of Childhood Cancer (ICC3) in cross-border data exchange.
5. ERNs should refer to the Guidance published by RD-ACTION codification experts, especially the **Tool-Kit** and the ***'Standard procedure and guide for the coding with Orphacodes'***

SECTION 2: CAPTURING PHENOTYPIC INFORMATION IN ERNs

i. In complex rare diseases, a patient may have the same (apparent) genetic mutation/anomaly but exhibit very different clinical presentations, with varying severity and prognosis. To capture and understand these variations, and translate this knowledge a) to better diagnostics and care for the patient under review, and b) to drive forwards the pace of knowledge and understanding for the field at large, it is often necessary to capture detailed phenotypic descriptions

ii. Given the scarcity and thus value of data in the rare disease and specialised healthcare field, it is important to optimise the utility of this clinical information, in terms of immediate, one-to-one patient benefit but also re-use, for instance by searching databases and computing similarity: the best way to do this is to use an agreed ontology for capturing phenotypes.

iii. The Human Phenotype Ontology or HPO is considered the most appropriate ontology for capturing the clinical presentation of rare diseases.[13]

iv. In recent years, efforts have been ongoing to harmonise clinical terms from the HPO with clinical symptoms recorded in the Orphanet database (i.e. the symptoms associated with a given disease) - 124,000 annotations of 7,700 diseases have been completed.

v. The HPO is used in various formats, within the diagnostics and care context:

a. Where clinicians or researchers agree that there are specific symptoms and presentations they wish to record for each patient seen in a consultation/followed-up in a study, a form-type interface can be very beneficial as clinicians are guided as to what items of information to provide and key elements are not omitted. In such circumstances, tools such as PhenoTips[14] can be very useful.

[11] ERNs or HCPs seeking hands-on guidance and support in adopting the ORDO might consider consulting Ontology experts or FAIR data experts.

[12] A tool-kit of resources to support these activities is available here – http://www.rd-action.eu/european-reference-networks-erns/rd-action-workshop2/ An official letter of invitation for collaboration were sent to ERN coordinators by the Orphanet Director in June 2017.

[13] Like the ORDO, the HPO has received 'IRDiRC Recommended Resources' status: http://www.irdirc.org/activities/irdirc-recognized-resources/

[14] https://phenotips.org

b. Often, clinicians prefer to capture information on their patients in a 'free' summary form, making open and unrestricted observations on the patient's presentation. Use of text-mining/auto-suggest software built upon the HPO now turns such text into an ontology-ready form, rendering it searchable and interoperable -type functionality. Systems such as the Patient Archive in Australia use open and very innovative software http://bio-lark.org/cr_restapi.html

c. In certain situations, a combination of a) and b) might be logical (for instance where one wishes to record the presence or absence of a number of compulsory clinical symptoms, but also leave space for an open and unrestricted body of clinical observation)

vi. Using HPO terms to construct clinical care records not only supports the interoperability of clinical data across ERNs, but also opens up this data to the vast repositories of interoperable data collected through research initiatives

vii. Efforts are underway to translate the HPO in 7 major European languages. This holds significant potential to enhance the clinical summaries captured in virtual 'review' or 'referrals' under an ERN.

RECOMMENDATIONS FOR THE CAPTURING OF PHENOTYPIC DESCRIPTIONS IN ERNs AND THEIR CONSTITUENT CENTRES

In view of the considerations above, the following are hereby recommended:

1. ERNs and their constituent HCPs should promote use of the HPO as the most appropriate ontology for capturing phenotypic descriptions in patients with a suspected rare disease or those requiring highly specialised procedures/techniques in which there is a need to build an evidence base.
2. ERNs -and particularly their common systems for exchanging patient data, such as the Clinical Patient Management System- should consider how best to use HPO[15], depending on the type of data collected:
   a. A free-text predictive tool powered by HPO is recommended when generating open clinical summaries and observations on the patient under review, to automatically create a structured phenotype profile
   b. If specific data items are being collected for research purposes, and/or the user wishes the record whether specific symptoms are present in a patient or not, a more prescriptive 'form-based´ system powered by HPO should be used
   c. In the context of virtual patient review/consultations, perhaps a combination would be best – the items that ERNs know they will always wish to monitor when reviewing a patient can be incorporated into a form structure, and the case report could also include a free text box for clinical notes.
3. When and where possible, ERN communities should seek to evaluate and improve the relevance of HPO terms in their particular thematic grouping/subdomain, by liaising with the HPO development team and considering the organisation of a dedicated workshop for these purposes.

[15] ERNs or HCPs seeking hands-on guidance and support in adopting the HPO might consider consulting Ontology experts or FAIR data experts. A good place to start is the Tool-Kit

SECTION 3: DEMOGRAPHIC DATA AND PATIENT PSEUDONYMISATION

i. Patients providing their data to an ERN for the purposes of receiving 'care' (i.e. those agreeing to a virtual referral) will be asked to sign an informed consent form. This form will also offer the patient the option to decide whether or not to consent to their data being retained in/by this Clinical Patient Management System, for reuse[16] (i.e. for a purpose beyond their own direct diagnosis, treatment and care). In either case, it appears that all patient data will be pseudonymised upon entry to the Clinical Patient Management System.

ii. As it will sometimes be necessary to discuss patient cases *across* ERNs, it is important that ERNs pseudonymise patients in the same way, to know that a given patient is one and the same person.

iii. Given the scarcity of data in the rare disease (and specialised healthcare) field, efforts were launched by the global research community to agree a common means[17] of pseudonymising patient data, to preserve privacy whilst affording researchers the opportunity to ascertain whether data held in myriad resources (e.g. registries, EHRs, biobanks, bioinformatics platforms, etc.) relates to the same patient. This mission gained prominence under RD-Connect[18] and is currently being advanced by a dedicated Task-Force, established under the IRDiRC (International Rare Disease Research Consortium) and the GA4GH (Global Alliance for Genomics and Health)[19].

iv. To allow the possibility of linking data from the same patient when collected and stored by different actors, it is necessary to either a) have a means of constructing an identifier in such a way that each patient will *always* receive the *same* identifier, no matter who requests that identifier, or b) to have a means of connecting different pseudonyms granted to the same patient (which involves the use of a trusted third party). Ideally, the most secure systems will exploit both direct identifiers (such as DOB) and quasi-identifiers (such as social security numbers)

v. The concept of a PPRL (Privacy Preserving Record Linkage) system has obtained favour in the global rare disease expert legal community -through the aforementioned Task-Force- in view of the potential to link records without knowing the identity of the individual, but with a high level of precision and recall.

vi. For ERNs to be able to participate in international rare diseases research efforts in future, they need to capture agreed elements of patient identifiable information in order to be able to pseudonymise the data in a particular way. Using an alternative means of pseudonymising patients will have far-reaching consequences, and could seriously hamper the ability of ERNs to achieve their research potential and in turn translate 'research' findings to better care for patients.

RECOMMENDATIONS ON PSEUDONYMISATION OF PATIENTS IN ERNs

In view of the considerations above, the following are hereby recommended:

[16] As per Descriptive Document for Tender SANTE/2016/A4/013 https://etendering.ted.europa.eu/document/document-file-download.html?docFileId=18496 2.5.1 (viii and ix)

[17] One example of such an approach - already influential in the European paediatric oncology community- is the EUPID, a privacy-preserving, secure and versatile system for pseudonymised patient registration and record linkage (https://eupid.eu/#/concept)

[18] http://rd-connect.eu/

[19] The Task-Force is preparing both ethico-legal recommendations and a technical solution.

1. ERNs should all collect the *same* core items of personally-identifiable demographic data (ideally including both direct identifiers and quasi-identifiers), in the same format, for each patient referred for shared care in the ERN (and ideally also, in time, for every patient seen by each HCP and potentially enrolled in registries or other essential resources), as a basis for generating an interoperable pseudonym for patients.
2. ERNs should embrace the state of the art in global efforts to agree a common means of pseudonymising patients with rare diseases and rare cancers, and embed the solution espoused by the international expert community (i.e. the dedicated Task-Force under IRDiRC and the GA4GH), namely, a Privacy Preserving Record Linkage (PPRL)
3. ERNs should agree their core demographic[20] items and corresponding formats –as per recommendation point 1, above- in accordance with the recommendations of the PPRL Task-Force under IRDiRC and the GA4GH
4. The system ultimately adopted by the ERNs to pseudonymize data should support a federated approach to ensure resilience of the system against a single point of failure and to mitigate risks of 'lock-in'.

SECTION 4: FAIR-IFYING DATA

i. As greater volumes of data are now collected and shared electronically, it becomes more feasible to exploit the full potential of that data. To achieve this, clinical information must be captured in a 'computable' form and made available through a controlled access mechanism. Against this backdrop, the concept of FAIR data is growing in prominence.[21] FAIR data principles[22] encourage robust management of data and metadata (i.e. data *about* data) for efficient use and reuse by humans and computers. FAIR principles prescribe that data is:

   ✓ Findable - (meta)data is uniquely and persistently identifiable and should have basic machine readable descriptive metadata.

   ✓ Accessible - data is reachable and accessible by humans and machines using standard formats and protocols, noting that 'accessible' does not equal 'open'

   ✓ Interoperable - (meta)data is machine readable and annotated with resolvable vocabularies/ontologies.

   ✓ Reusable - (meta)data is sufficiently well-described to allow (semi)automated integration with other compatible data sources.

   (See footnote 22 for the full reference description of the FAIR guiding principles)

ii. FAIR principles are intended to support data consumers in efficiently querying and analysing data created by different actors, to achieve a defined health-related or research goal. When source data are sensitive, FAIR principles prescribe that non-sensitive information can be *virtually aggregated* from these sources without ever *revealing* the sensitive data: FAIR-compliant software services control access, to remain within legal and ethical boundaries.

[20] To be confirmed in the forthcoming recommendations of the aforementioned Task-Force, but likely to include (for rare diseases) items such as First name, surname (as on birth certificate), Date of birth; and perhaps also middle name and city of birth as on birth certificate.

[21] Organisations that endorse FAIR data principles include ELIXIR, BBMRI, the European Open Science Cloud, FORCE11, NIH through its 'commons' program, and the G20.

[22] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016) https://www.ncbi.nlm.nih.gov/pubmed/26978244

iii. The preferred approach is to prepare data 'at source' for broader use and analysis across different resources.

iv. An important element of FAIR-ifying data in rare diseases is the use of standardised, consensus ontologies, as addressed above for diseases (section 1) and phenotypes (section 2). A FAIR-ification process finds appropriate ontologies for all data at-hand (ontologies exist for most types of data in life science and medicine).

v. It is possible to 'FAIR-ify' data *not* collected in a FAIR manner initially: to make such data interoperable and machine-readable, it is necessary to explicitly define appropriate ontological terms and relationships between data items retroactively. Access to interoperable and original data can be controlled via an API (Application Programming Interface). One such API is the FAIR Data Point, which uses the Data Catalogue Vocabulary [23] to specify the metadata of a data source in a standard, machine-readable format.[24] APIs are typically used in user-oriented software applications or in computational analysis workflows.

vi. Recommendations/good principles regarding the coding of diseases and phenotypes, pseudonymisation, privacy preservation, and consent are essential for a FAIR approach, as all constitute aspects of FAIR data stewardship.

RECOMMENDATIONS ON INCORPORATING FAIR DATA PRINCIPLES TO ERNs

In view of the considerations above, the following are hereby recommended:

1. ERNs should consult rare disease FAIR data linkage specialists to discuss their specific needs and opportunities to link clinical data generated by ERNs -and ideally their constituent centres- with additional data sources.
2. The most logical point of engagement is the new GO-FAIR[25] implementation network currently being established in the rare disease domain, the main goal of which is to professionalize FAIR services for rare diseases and ERNs.
3. ERNs and their constituent HCPs should stimulate working on local data quality via FAIRification, ideally in consultation with FAIR data experts:
    a. Where possible, stakeholders should consider organising a dedicated FAIRification project
    b. If a full-scale project is not feasible (or not feasible at present), ERN data experts are advised to participate in a FAIR-led 'Bring Your Own Data' workshop, defining at least one 'driving user question' to address.
4. When writing grants or planning activities which pertain to the generation, processing or exchange of data, ERNs should include a data management plan espousing FAIR guiding principles.[26]
5. When dealing with valuable data describing neither diseases nor phenotypes[27], it may be advisable to consult ontology and FAIR data experts regarding the application of terms from appropriate ontologies and how to link these correctly to HPO and ORDO terms (if and when appropriate).

[23] https://www.w3.org/TR/vocab-dcat/

[24] https://www.researchgate.net/publication/309468587_FAIR_Data_Points_Supporting_Big_Data_Interoperability

[25] https://www.dtls.nl/fair-data/go-fair/

[26] For instance by including a requirement for any software service providers to present their plans towards compliance with FAIR guiding principles

[27] Addressed in Sections 1 and 2, respectively.