H2020-MSCA-ITN-2018-813545

HELICAL

Health Data Linkage for Clinical Benefit

Deliverable 29 (6.7): Open Data research Pilot

This deliverable reflects only the authors' views, and the European Commission Research Executive Agency is not responsible for any use that may be made of the information it contains.

This project has received funding from the European Union's Horizon 2020 research and Innovation programme under the Marie Skłodowska-Curie grant agreement No 813545

Table of contents

1. Introduction

European researchers have made leading contributions to the large genomic, transcriptomic and clinical datasets from patients with chronic diseases. As part of the HELICAL programme, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813545, Early Stage Researchers (ESRs) have conducted an analysis of large datasets, using autoimmune vasculitis as a paradigm, through the application of informatics to these datasets in order to gain new biological insights.

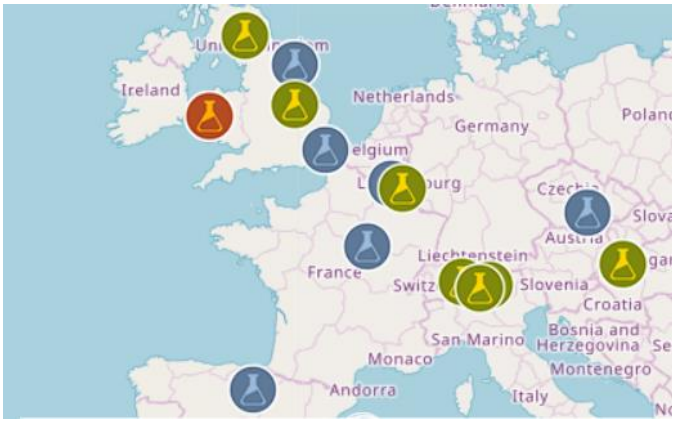


Figure 1: HELICAL partner sites' locations

The ESRs process and derive (i.e. use and generate) research data and metadata as well as software and relevant algorithms as part of their projects. Research data includes comprehensive biological and clinical datasets collected from patients with a rare condition. For example, data about the location, age, sex and images showing the kidney tissue of a patient. Research metadata consists of dataset descriptors together with information about the origin and processing of the data. The software describes the routines conducted during the data processing.

Given the particular sensitive nature of the research data gathered, the researchers have proceeded, where appropriate and possible, to pseudonymise the datasets in order to be in compliance with the General Data Protection Regulation 2016/679 (GDPR) as well as other national data protection laws and regulations. Further clarification on whether all ESRs have access to this type of data, whether sites have pseudonymised the data or some datasets have been anonymised is required.

The aim of the HELICAL Open Research Guidelines document is to present a set of recommendations and step-by-step instructions to help ESRs and experts share their findings (whether aggregated data, software or algorithms) as open as possible in accordance with the relevant data protection rules and legislation.

More specifically, the Guidelines aim to provide an understanding and exemplification as to how research findings (data and analysis) can be published securely in an open-access manner. Further details as to the content of the Guidelines is described in the sections below.

2. Objectives

The following section describes the objectives to allow ESRs to publish their data in an open access manner.

2.1. Assembling HELICAL data types

A Google Docs spreadsheet is used to gather information about the data types used in each of the ESR projects. This approach promotes the understanding for the overall HELICAL data types as well as the specifics for each ESR.

Progress against the objective

The ESRs recorded the data type used in their individual projects as a Google Docs spreadsheet shared internally for project use. A copy of the spreadsheet is included in **Annex 1**. The spreadsheet provides columns for demographic, clinical, biomarkers, genetic, experimental and environmental data with appropriate subcategories for a better understanding of the data types. In addition, the spreadsheet was extended to include a data analysis column reporting the methods and data combinations used, a publish data and code column to detail the requirements for each ESR in publishing and a data example column simplifying the understanding of the data used by the ESR. The spreadsheet information was then reviewed by Maria (ESR15), Bahareh (ESR3), Filippo (ESR9) and Albert (ESR1) to validate the imputed information for each ESR. The validation process involved arranging individual calls with the ESRs to clarify and complete their inputs in the Google Docs spreadsheet.

2.2. Identifying EU platform to publish HELICAL data

Identifying a EU platform to publish HELICAL data which is used, generated or gathered from the ESRs would facilitate, guide and enhance the publication data process for the project.

Progress against the objective

The Open Research Europe platform[1] was identified as the most appropriate platform to publish HELICAL data. The EU platform “allows Horizon 2020 beneficiaries to comply with the open access terms of their funding” which would offer the ESRs a venue to share their results and insights in a rapid and compliant manner which would enable open, constructive research discussion. Through a reading of the eligibility criteria, the HELICAL project has been identified to meet the threshold for publishing in the Open Research Europe platform. Furthermore, publication costs are covered by the European Commission waiving the fees for the HELICAL project. The platform has an extended documentation section providing guidelines to publish data from a practitioner perspective[2].

2.3. Designing guidelines for HELICAL ESRs to publish their data

The guidelines are a series of steps to support the ESRs in publishing their data in the Open Research Europe platform.

Progress against the objective

An initial guideline for the ESRs has been designed by the authors of this document. The guideline comprises the steps from identifying existing literature examples with similar research to deposit the data in the Open Research Europe platform. Furthermore, the guideline encourages the ESRs to check standard terms used for their data variables facilitating data reuse within the HELICAL Work Packages (WPs).

The consensus regarding the terms for each WP will be reviewed by a WP representative. In addition, Maria (part of WP4) will support in addressing any queries on data protection issues arising from the datasets and analysis methods, and the code licence questions which may arise from working alongside industry ESRs, thereby facilitating the comprehension of the guidelines for the researchers.

[1] European Commission, Directorate-General for DG RTD, ‘Open Research Europe’ open access publishing platform. Available at: https://open-research-europe.ec.europa.eu/

[2] Ibid, Data Guidelines. Accessed on 1st March 2022. Available at: https://open-research-europe.ec.europa.eu/for-authors/data-guidelines

3. The Guidelines

A draft guideline was designed by the authors after meeting every two weeks for an online meeting for the 6 months. The draft guideline to facilitate publishing data on an open access repository for the ESRs is structured as (i) a set of prerequisites that need to be met prior to proceeding to publication and (ii) a series of steps to assist in open data publication.

3.1. Open data prerequisites

Prerequisite 1: Gaining/Granting Permissions

When making data available as open data in an open access repository, it is presumed that the processing conducted and datasets used have already been approved as compliant by your institute, company, data owner and/or HELICAL's Data Protection Officer (DPO).[3] In this context it is important to note that (1) *gaining* permission to use data and (2) *granting* permission to another to access that data are two very different concepts.

Gaining permission to use the data

ESRs should have already conducted a Data Protection Impact Assessment ('DPIA') providing information as to the legal basis, processing activities (e.g. data analysis, software used, data havens) or sufficient permission that has been granted to the relevant data custodian (e.g. registry and/or databank)[4] by the data subject[5] (in this instance a patient or person with a rare condition) to access and use the data. Further information as to the drafting of a DPIA were provided during the delivery of Modules 2 and 3 'Ethical linking of electronic health data to research data to support research, Open Science and uphold FAIR principles'. The Module documents are included on Basecamp and the tailored HELICAL DPIA template can be found in **Annex 2**. The following diagram in Figure 2 exemplifies the overall way the data flow beginning from the patient and ending with the researcher.

[3] In accordance with the legal basis under the GDPR, accompanying recitals, guidelines issued by the European Data Protection Board and other data protection laws and regulations.

[4] The term 'custodian' is not defined under the GDPR but refers to someone that has administrative control of a document or electronic file. The closest defined term under the GDPR would be that of a "data controller" as defined under Art. 4(7).

[5] A data subject, as per Art. 4(1) GDPR, refers to "an identified or identifiable natural person".

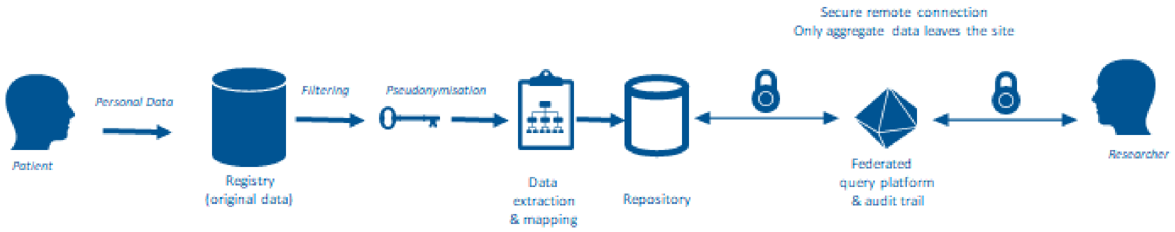


Figure 2: Diagram of data flow

Granting permission to use the data

Publishing the data as open data equates to granting permission for the public to access and use the data themselves. Therefore in order for this to be permissible, **the data subject needs to have consented to their data being published, and the data controller should have been granted the relevant permission to pass on and publish the data in an open access repository**. The ESRs should be aware of the relevant permissions and reflect these in the project's DPIA if they would like to move to publication of the data as Open Data.

The ESRs in this regard should consider: (i) whether the legal basis or permission covers all of the data which will be published as part of the project; or (ii) whether this applies only to parts of the dataset. If the latter, the ESR should create a subset or an example[6] from the data that can be published as open data.

Given that the majority of ESRs use 'personal data' which can be considered as 'sensitive' (according to the definitions in Articles 6 and 9 of GDPR), these already would be protected through means of anonymisation or pseudonymisation[7]. However, the rare disease nature of the data presents an obstacle for the anonymisation process. Effective anonymisation might not be possible without losing value for research due to the low sample sizes and specialised type of the data (e.g. genomic, genetic and image data). **Each ESR is responsible for providing enough information to the data controller and DPO for them to assess if the data or which data is ready to be published as open data.** Further guidance on anonymisation can be found in the materials and recording of the sessions given as part of the delivery of Modules 2 and 3 of the ESR HELICAL training in June 2020 available on Basecamp.

[6] Example data means simulated data which follows the same structure as the real data for reproducibility.
[7] Some ESRs will collect new personal data during their data collection process. In this case, they should receive approval by an Ethics Committee and be able to make the dataset anonymous prior to publication.

When the data is ready for publication as open data, **data will be accompanied with sufficient research metadata so other researchers can reuse it in different contexts**. In addition, the metadata used in the HELICAL project should use common vocabularies to guarantee collaboration within the consortium. In particular, **metadata should include** a detailed explanation of the following fields: data variables (e.g. name, definition and units), dataset descriptors (e.g. title, publisher and licence), processing purpose, origin and processing of the data (e.g. datasets used and statistical methods applied), hereafter data provenance and lineage.

In addition, if ESRs have permission to publish the data in an open access repository, they need to ensure that the data being published meets the FAIR principles. For this to be achieved, the data should meet the four foundational principles—Findability, Accessibility, Interoperability, and Reusability - to the greatest extent possible, as per Article 5 GDPR. **ESRs should be able to provide evidence as to how the FAIR principles have been met and implemented for the research datasets that are to be published.** This will differ based on "the data types, the nature of research (e.g. ethical sensitivities or commercial partners) and the level of existing support for data sharing"[8]. The FAIR principles can be applied to the design stage of the research to facilitate the documentation of the implementation process followed through the updating of the DPIA.

<u>It is stressed that ESRs cannot proceed on the open data pathway if the 'Gaining/granting permission' prerequisite is not met.</u>

'*In that context, the implementation of FAIR data needs to go hand-in-hand with the principle that data created by publicly-funded research must be as Open as possible and as closed as necessary.*
The FAIR principles - and related concepts and policies - should be applied not just to data, but to metadata, identifiers, software and Data Management Plans (DMPs) that enable data to be FAIR.'

- European Commission, 'Final Report and Action Plan from the European Commission Expert Group on FAIR Data: Turning FAIR into Reality', Directorate-General Research and Innovation, 2018

[8] European Commission, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 'Turning FAIR into Reality' Directorate General for Research and Innovation, 2018. Available at: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_0.pdf

Prerequisite 2: Ensure Traceability

ESRs should keep a track record of the processing steps applied to the initial data input. The documentation for the processing steps **should include** the timeframe, the data controllers and processors, and the data lifecycle (ordered annotated amendments with the corresponding version, data analysis steps, data combinations, aggregates and subsets) conducted on the data.

Sufficient metadata (as defined in *Prerequisite 1*), included in the dataset publication, can cover most of the processing steps documented. However, we encourage the ESRs to provide notebooks (e.g. Jupyter notebooks or Markdown reports) for the data analysis conducted at the time of publication.

This is a mentality that the ESRs need to keep in mind from the beginning of obtaining the data to the point of publication. We note that this documentation should also reflect any steps applied prior to the initial data input. This in return demonstrates a culture of traceability for other researchers to make trustworthy valid inferences from the published data.

It is stressed that ERSs cannot proceed on the open data pathway if the ‘Ensure Traceability' prerequisite is not met.

3.2. Steps towards publication

Step 1: Record data type used in the Google Doc spreadsheet

The ESRs are responsible for recording the data type used in their individual projects as represented in the provided Google Doc spreadsheet. The recording of all data types would support the scientific workflow of the ESRs providing the datasets used in intermediate steps. Once this is finalised, this will require confirmation by the individual supervisors of each project and each change or amendment will need to be reflected in the individual DPIA by the ESRs.

Step 2: Provide an example representing the data collection and analysis processes

Each ESR is tasked with researching and providing an example from literature, which would match their individual project and utilisation of data, algorithm and/or software. This example has already been published in an open access manner to clarify visually the type of data that is used. By means of example, the following footnote provides a notebook describing the data analysis steps followed for an aspect of the project conducted by ESR6. A copy of this can be found under **Annex 3** and at the following link:

- Alejandro Fontal. (2022). helical-itn/data-usage-examples: 0.0.1 (v0.0.1). Zenodo. https://doi.org/10.5281/zenodo.6340349

Step 3: Agree on a standard data vocabulary

Having had initial discussions with the ESRs and by reviewing the populated Google Doc spreadsheet, inconsistencies in the naming of terms representing seemingly either highly similar or identical values are named differently in each project.

In order to ensure consistency, and to create a consensus in every WP, the creation of a common understanding about what each of those terms correspond to and the creation of a common or standard data dictionary for the variables used by the ESRs in each WP should be established. To achieve this the WPs representatives will communicate with the remaining ESRs and through discussion will come to an understanding and agreement as to what the variables in every project in their WP represent and put together a standard data dictionary. Researchers should ensure that a standard data dictionary has not been already agreed to.

Step 4: Consult existing guidelines from individual research sites

Every university, company and research institute has access to the guidelines concerning the data sharing of their research data and analysis. ESRs should research for, identify and communicate the guidelines to the HELICAL Open Research Guidelines team, composed by the authors Maria and Albert. The team will then review the Guidelines to ensure that any subsequent steps or suggestions included in the document are in line. The Guidelines should be accessible to the ESRs and should already be reflected in their DPIAs. The ESRs must make sure that there is a consensus on the content of these internal guidelines and the Information Governance Policy of the HELICAL project, accessible through Basecamp. A copy of the policy can be found under **HELICAL Information Governance Policy - Deliverable 4.2**.

Step 5: Familiarise with the FAIR data principles

All ESRs need to become familiar with the FAIR principles in order to ensure that their projects are compliant. In this regard, the ESRs have been provided with various educational materials for this purpose as well as throughout the course of the HELICAL project. Examples of these include:

I. The materials shared and presented during the delivery of Module 2 in April 2020, specifically sessions 4 and 5 - materials of which are stored under the Training folder of the Full Consortium file, accessible on Basecamp.

II. The FAIR data principles[9] and the FAIRy tale[10] from Zenodo.

III. The open data platform article[11] and data guidelines.[12]

Step 6: Publish research data in a open data repository

We recommend publishing your research data in the open data repository preferred by each community. Different communities might prefer one repository from another based on the flexibility of the licence, the fields in the description, the collaboration with particular journals, the user interface or the data upload limit per publication. Researchers not clear on how to choose the data repository preferred by the community can check nature data repositories[13] and re3data[14] to inform their decision.

However, we wanted to share a general go to open data repository to have a starting point:

Zenodo – https://zenodo.org/ and Github – https://github.com/

Zenodo allows researchers to publish data in any size and format, which addresses the diverse data types and formats within the project (e.g. clinical data, environmental data, image biopsies or questionnaire data). In addition, researchers can describe their data so it is easier to reuse by others, which can then be linked to related research (e.g. journal publications). When publishing the data in Zenodo, the content will automatically obtain a Digital Object Identifier (DOI), making the content citable and discoverable.

The ESRs can link their publication with the ITN HELICAL community to have a central repository to access HELICAL open source artefacts. The steps to link the publication are the following: (1) log in to your Zenodo account, (2) choose the publication to link, (3) click on edit (top right corner), (4) select ITN HELICAL Community, (5) save and publish. The linkage request will then be approved by the curator of the community.

Github cloud-based hosting service that allows developers and researchers to store, manage their code, while tracking the changes to their code during the developing process. Furthermore, Github code can be linked to Zenodo which will generate a DOI

[9] Zenodo, Principles. Accessed September 2021. Available at: https://about.zenodo.org/principles/
[10] Zenodo, A FAIRy tale. Accessed September 2021. Available at: https://zenodo.org/record/2248200#.YG85UBLTVhG
[11] Zenodo, Guidelines for authors. Accessed September 2021. Available at: https://open-research-europe.ec.europa.eu/for-authors/article-guidelines/
[12] Ut supra 8. Available at: https://open-research-europe.ec.europa.eu/for-authors/data-guidelines
[13] https://www.nature.com/sdata/policies/repositories
[14] https://www.re3data.org/

for the code, making it citable. Below there are two examples of data and code published by ESR1 in Zenodo and Github:

1. Navarro-Gallinad, Albert. (2022). Environmental data associated to particular health events example dataset (20220713) [Data set]. Zenodo https://zenodo.org/record/6828075
2. Navarro-Gallinad, Albert. (2021). navarral/serdif-api: Release version 1.0.1 (v1.0.1). Zenodo. https://doi.org/10.5281/zenodo.5776245

However, prior to publishing the data accompanied with metadata and code to Zenodo and Github, the ESRs must get approval from their supervisors, the HELICAL team and the data owner that it can be published.

4. Acknowledgements

Thank you to all the ESRs for their help in aiding the collection of the required information to draft these guidelines and for providing their DPIAs in order for the present document to correctly reflect the steps followed in practice.

Particular thanks to Bahareh Kosravi and Filippo Guerri in their participation and contributions during the first meetings that were held in late 2020 to coordinate efforts as well as for their help in reviewing and organising the information inserted in the Excel spreadsheet containing the data used by the ESRs in the project.

5. Annexes

The Guideline will include an additional annex to track the publication process for data, code and workflows for each of the ESRs. The annex will be structured with the steps from the presented guidelines.

Annex 1: HELICAL ESR data types

ESR	Data type	Data subtype	Description
	Demographic	Individual	
		Population	
	Clinical	Electronic Health Records (EHR)	
		Questionnaire	
		Images	
	Biomarkers	Serologic	
		Inflammatory	
		Others	
	Genetic	Gene expression	
		Sequencing	
	Experimental	Collected	
		Shared	
	Environmental	Weather	
		Air pollution	
		Aerosol	
		Soil	
	Data use processing	Data analysis	

		Datasets	
	Publish	Dataset	
		Code	
		Workflow	
		Example	
	Open category	Other	

Annex 2: HELICAL tailored DPIA

IG Assessment Checklist – [Project Title]

Introduction to IG Assessment process

Under the General Data Protection Regulation (GDPR), a Data Protection Impact Assessment (DPIA) is only required where proposed data processing is “likely to result in a high risk to the rights and freedoms of natural persons” (Article 35(1)). However, Article 35(3) explicitly requires one where there is ‘large-scale’ processing of ‘special category’ (e.g. healthcare) data then a DPIA is required.

One other possibility is that the data being processed is already anonymised (see Recital 26) so falls outside GDPR altogether so that no DPIA is actually required.

However, good project management and information governance suggests that there should be a general approach to risk assessment for any project or business enterprise – if only to determine whether a DPIA might be required.

Ideally, one should work from a simple initial Checklist (this document) which identifies possible areas of information risk and compliance requirements to a ‘discussion note’ which explores any issues in more depth and may help identify the necessary mitigation methods and mechanisms to offset most if not all risks. Only if risks are unmitigated or remain ‘high’ would you move to a formal DPIA report.

The IG Assessment approach

There should be an overview of the proposed project or business change to explain what processing is envisaged as well as the purpose and intended outcome. The ‘purpose’ is important to establish the legal basis for the processing as well as ensuring that any possible mitigations or counter-measures do not undermine the main rationale for the processing.

The next step is to establish what compliance requirements may apply: GDPR, contractual or other regulatory restrictions, consent requirements, or obligations to preserve the data for legal or other reasons (including the benefit of posterity perhaps).

Once the precise range of obligations has been established, then appropriate checks can be made and recorded within the document.

The most obvious of these being GDPR compliance. There must be a ‘High Risk’ assessment (Appendix A) to determine whether the supervisory authority needs to be informed – generally, it is expected that it will not be necessary; if so, then a formal DPIA report will be needed.

Appendix B has a broader Privacy Impact Assessment that may throw up some broader issues.

Initial conclusions as to next steps or particular countermeasures to be considered should be detailed below.

Project Background/Overview

[Explain business background, including any existing processes and procedures; outline the project including stages, deliverables, and timelines]

Comparison of process steps (simplified): [optional]

This allows identification of what processing is new or changed through the project:

Step	Current	Proposed
Project initiation, including any ISAC approval, up to Task Order from client		No change

Initial Conclusions

concerning further counter-measures or business viability [possibly tentative]

1. ...
2. …

Compliance Checks required:

Tick	**Requirement**	**Notes** [replace guide text with response]
□	Does the project involve processing ‘personal data’ of any sort?	Note: not just patient data; may need clear assessment of any anonymization to establish outside GDPR
□	Does the project involve processing ‘confidential data’ of any sort?	Note: may be ‘commercial in confidence’, medical confidentiality, or organisational confidentiality (internally sensitive); may need to check contractual limitations
Data Availability requirements		

□	Does data need to be held for GCP compliance?	
□	Does data need to be held to meet ‘Open Data’ requirements?	
□	Does data need to be held to meet ICMJE requirements or commitments?	

GDPR Compliance Checklist – where ‘personal data’ is processed:

Tick	**Requirement**	**Notes** [replace guide text with response]
Article 5: Principles compliance checks		
□	a) Is processing lawful, fair, and transparent?	
□	b) Is the purpose (or purposes) of the processing clearly defined	[‘purpose limitation’ so should cover any subsequent or later processing]
□	c) adequate, relevant and limited to what is necessary	[‘data minimisation’]
□	d) accurate and, where necessary, kept up to date	
□	e) kept and permits identification of data subjects for no longer than is necessary	[‘storage limitation’]
□	f) processed securely	
□	2) can you demonstrate this compliance?	[‘accountability’]
Articles 13 & 14 compliance		[See detailed Transparency Checklist below]
□	Did the data came from publicly accessible sources?	[if so then transparency requirements may be reduced, but need to ensure data is accurate & up-to-date]
□	Are data subjects informed before processing starts for any new purpose if incompatible with original purpose where the controller wants to use data	

	for a different purpose to the purpose for which they currently hold data	
□	Does the Privacy Notice and/or PIL cover this processing?	
□	What patient choices are available? Are these explained?	[see also Data Subject Rights below]
Articles 6 and 9: legal bases		
□	What are legal bases under Article 6	
□	What are legal bases under Article 9 (if 'special category' data)	
□	Are Article 6 legitimate interests explained where relevant?	[Complete an LIA form]
□	Are details of statutory obligations for Article 6 explained where relevant.	[Quote statutes or regulation]
□	Is this proposed processing compatible with the declared purposes?	[Check against any privacy notices and public information]
Article 89(1) research exemption		
□	If for research, do we meet Art 89(1) data minimisation	
Articles 15-23: Data Subject Rights		[See detailed table below]
□	Do we support data subject rights?	[If data is pseudo-/anonymised, then it would be difficult/impossible to do so]
□	There is no use of automated decision making (e.g. profiling)	[Otherwise need at least a 'discussion note']
Articles 24-43: Controller-Processor		
□	A28 & 29: What measures are there to ensure processors comply?	[Is there a formal Data Processing Agreement]
□	A30: Is there an entry for this processing/data held in the register?	
□	A32-34: Do we ensure appropriate security, including protection against unauthorised or unlawful processing and against accidental loss,	[separate security checklist?]

	destruction or damage, using appropriate technical or organisational measures?	
□	A37-39: Is there a DPO and have they been or will they be consulted?	[part of sign-off of the DPIA]
Articles 44-50: International transfers		
	What form of data will be transferred to a third country or international organisation	[describe nature of data and whether identified, identifiable, de-identified or anonymous]
□	Are there safeguards for international transfers?	[e.g. US Privacy Shield, anonymisation, GDPR equivalence, approved contractual clauses, or BCR]
Article 90: Obligations of secrecy		
□	Do we meet medical confidentiality requirements?	[Note any national case law and statutory requirements that may affect this]

Data Subject Rights:

Note if supported and what process/procedure applies; if not, then describe the legal justification for not supporting this right.

□	To be informed: about processing, about choices, about rights, about controller	
□	the right of access to see or receive a printed copy	
□	the right to rectification – to correct any material errors in the personal data	
□	the right to erasure – where appropriate, to ask that all personal data is erased	
□	the right to restrict processing – to ask that some or all processing ceases [see opt-out]	
□	the right to data portability – this only applies to data provided directly by individual	
□	the right to object to and not to be subject to automated decision-making, including profiling	

□	Right to object to a Data Processing Authority (typically the relevant supervisory authority of each Member State)	
□	Where consent is the legal basis, the right to withdraw consent	

Detailed Transparency Checklist[15]

Does privacy information provided to data subjects include:

□	The name and contact details of our organisation	
□	The name and contact details of our representative (if applicable)	
□	The contact details of our data protection officer (if applicable)	
□	The purposes of the processing	
□	The lawful bases for the processing	[Art6 for 'personal data' & Art9 for 'special category'
□	The legitimate interests for the processing (if applicable)	
□	The categories of personal data obtained (if the personal data is not obtained from the individual it relates to)	[for Art14]
□	The recipients or categories of recipients of the personal data	
□	The details of transfers of the personal data to any third countries or international organisations (if applicable)	
□	The retention periods for the personal data.	
□	The rights available to individuals in respect of the processing	
□	The right to withdraw consent (if applicable)	

[15] Taken from UK Information Commissioner's Office template

□	The right to lodge a complaint with a supervisory authority	
□	The source of the personal data (if the personal data is not obtained from the individual it relates to)	[For Art14]
□	The details of whether individuals are under a statutory or contractual obligation to provide the personal data (if applicable, and if the personal data is collected from the individual it relates to)	
□	The details of the existence of automated decision-making, including profiling (if applicable)	
□	We provide individuals with privacy information at the time we collect their personal data from them – or where e obtain personal data from a source other than the individual it relates to, we provide them with privacy information	
□	within a reasonable of period of obtaining the personal data and no later than one month	
□	if we plan to communicate with the individual, at the latest, when the first communication takes place	
□	if we plan to disclose the data to someone else, at the latest, when the data is disclosed	

□	We provide the information in a way that is:
	☐ concise;
	☐ transparent;
	☐ intelligible;
	☐ easily accessible; and
	☐ uses clear and plain language.
		[Describe how we check is Plain English, etc.]

□	When drafting the information, we:	[Note: best practice advice]

	☐ undertake an information audit to find out what personal data we hold and what we do with it. ☐ put ourselves in the position of the people we’re collecting information about. ☐ carry out user testing to evaluate how effective our privacy information is	

□	When providing our privacy information to individuals, we use a combination of appropriate techniques, such as:
	☐ a layered approach;
	☐ dashboards;
	☐ just-in-time notices;
	☐ icons; and
	☐ mobile and smart device functionalities.
		[Note: best practice advice]

Security & Access Control Checklist

Controls need to be appropriate to the level of risk: identified special category data needs more protection against potential misuse than non-personal data.

	Data Security classification (above Official)	□ - Official-Sensitive □ - Secret □ - Top Secret □ - Public Domain
□	Personal Data involved [GDPR]	
□	Special Category of personal data involved [GDPR]	
□	Electronic Communications (inc. cookies) [PECR]	
□	Credit Card data	
□	Legal enforcement [LED2018]	

□	Financial data	
□	Intellectual Property (detail owner)	
□	Commercial in confidence (detail owner)	
	Data Location (storage or processing) (include any back-up site(s))	□ - UK □ - EU/EEA □ - EU White-list □ - USA □ - Other:
□	Is data held in a secure data centre?	[detail centre and what certification supports assertion]
□	Is this new supplier, location, or system?	[If so, need specific IS check; also need formal contract]
□	Is all user access subject to 2-factor authentication?	□ - no control □ - single factor (e.g. just password) □ - 2-factor (e.g. password & fob) □ - biometric [note: GDPR reqs] □ - Other control:
□	Are there established JML procedures?	[Joiners, Movers, Leavers]
□	Are there checks that passwords are robust and secure enough?	[]
□	Are all administrator & user accounts routinely monitored?	[Particularly for redundant or little used accounts]
□	Are systems protected against malware and other attacks?	[provide details of protection software and procedures

[Need some aspect of CIA/impact-likelihood assessment]

Information Asset Register Checklist

□	Are there new IAs being created?	[provide details]
□	Are old IAs being retired?	[provide details]
□	Have IAOs & IACs been consulted?	

□	Has IAR been updated/amended?	[at least create project task to do so]
□	Data Retention classification & period	
□	Data retention procedure/functionality in place	

Appendix A – Supervisory Authority 'High Risk' Check

If the DPIA shows 'high risk' processing which cannot be mitigated, then the DPIA should be sent to the relevant authority for review before any processing starts. Note that their review may take several weeks to process. A 'High Risk' assessment represents a 'risk to the rights and freedoms of individuals' – so may extend beyond GDPR consideration, including Human Rights.

GDPR Article 35(3) provides three examples:

a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;

b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 1013; or

c) a systematic monitoring of a publicly accessible area on a large scale

ICO cites:

1. Systematic and extensive profiling with significant effects
2. Large scale use of sensitive data [viz. 'special category' in GDPR terms]
3. Public monitoring

These being the same as (a)-(c) above. They further identify:

1. **New technologies**: processing involving the use of new technologies, or the novel application of existing technologies (including AI).
2. **Denial of service**: Decisions about an individual's access to a product, service, opportunity or benefit which is based to any extent on automated decision-making (including profiling) or involves the processing of special category data.
3. **Large-scale profiling**: any profiling of individuals on a large scale.
4. **Biometrics**: any processing of biometric data.
5. **Genetic data**: any processing of genetic data, other than that processed by an individual GP or health professional for the provision of health care direct to the data subject.
6. **Data matching**: combining, comparing or matching personal data obtained from multiple sources.

7. **Invisible processing**: processing of personal data that has not been obtained direct from the data subject in circumstances where the controller considers that compliance with Article 14 would prove impossible or involve disproportionate effort.

8. **Tracking**: processing which involves tracking an individual's geolocation or behaviour, including but not limited to the online environment.

9. **Targeting of children or other vulnerable individuals**: The use of the personal data of children or other vulnerable individuals for marketing purposes, profiling or other automated decision-making, or if you intend to offer online services directly to children.

10. **Risk of physical harm**: Where the processing is of such a nature that a personal data breach could jeopardise the [physical] health or safety of individuals.

'High Risk' assessment using ICO criteria:

Criterion:	**Assessment**	**Comments**
New technologies		
Denial of service		
Large-scale profiling		
Biometrics		
Genetic data		
Data matching		

Invisible processing		
Tracking		
Targeting of children or other vulnerable individuals		
Risk of physical harm		

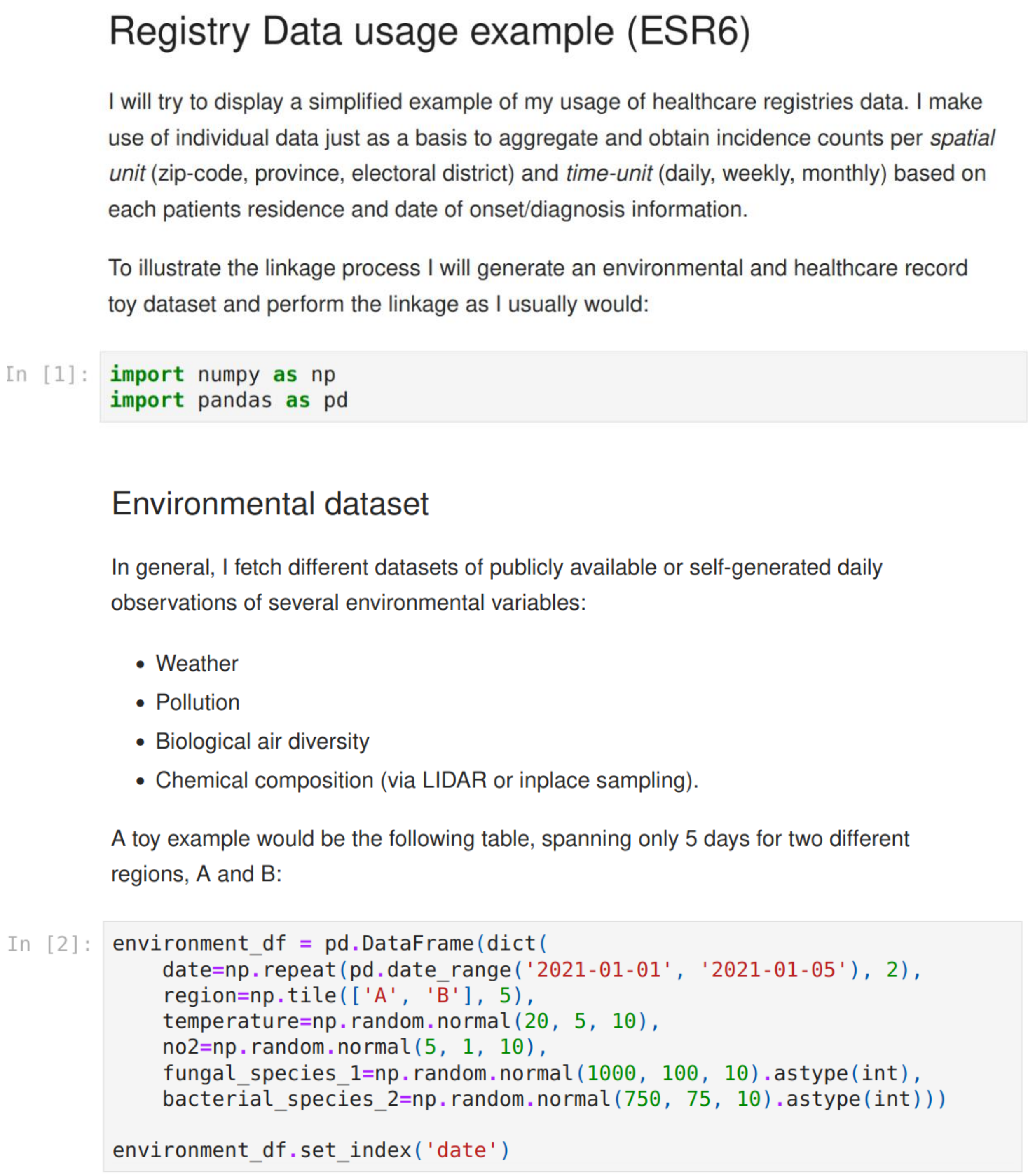
[The assessment can be one of N/A (not applicable), Low, Medium, or High. The comments should explain how the assessment is justified.]

Appendix B – Broad Privacy Risk Assessment:

#	Risk Description/detail	Discussion
1.	Data accuracy and timeliness	[Is data accurately recorded & kept up-to-date?]
2.	Differential treatment of patients/data subjects	[Might certain categories of people be adversely affected, e.g. children, vulnerable adults]
3.	Data Accuracy and identification	[Is the identification of individual reliable? Is there a danger of mis-attribution or incorrect linkage of data?]
4.	Holding / sharing / use of excessive data within i~HD systems	[Might too much data be held or for long? Is there a clear justification for data retention? Not 'just in case']
5.	Data held too long within i~HD systems	[Is there a clear data retention period specified and are there processes to ensure its deletion when no longer needed? Are copies tracked and deleted as well?]
6.	Excessive range of access in terms of users to personal data (consider new users/change of access privileges)	[Do more users have access than strictly necessary? Are user roles clear distinguished and reflected in the access privileges? Is there a clear process for granting and revoking access privileges?]
7.	Potential for misuse of data, unauthorised access to systems	[What are the likely threats to the data? What countermeasures are or might be applied? Is it possible for access to be granted inappropriately?]
8.	New sharing of data with other organisations, including new or change of suppliers	[Is data being shared from new data providers or with new data users? Are there new suppliers or data processors? What controls will apply?]
9.	Variable and inconsistent adoption / implementation	[How well will this system work end-to-end? How robust is it against partial adoption or system failure?]

10.	Legal compliance, particularly DP transparency requirements and support for data subject rights	[How well does this system meet legal requirements – or appear to meet legal requirements? Does it meet the ‘No surprises’ rule? What would happen if an individual requests data erasure or ceasing processing, etc.]
11.	Medical confidentiality	[Are there any addition sensitivities over confidentiality? Might specific approval (e.g. REC) be required to support this processing?]

Annex 3: Exemplary notebook describing data analysis steps



Registry Data usage example (ESR6)

I will try to display a simplified example of my usage of healthcare registries data. I make use of individual data just as a basis to aggregate and obtain incidence counts per *spatial unit* (zip-code, province, electoral district) and *time-unit* (daily, weekly, monthly) based on each patients residence and date of onset/diagnosis information.

To illustrate the linkage process I will generate an environmental and healthcare record toy dataset and perform the linkage as I usually would:

In [1]:
```
import numpy as np
import pandas as pd
```

Environmental dataset

In general, I fetch different datasets of publicly available or self-generated daily observations of several environmental variables:

- Weather
- Pollution
- Biological air diversity
- Chemical composition (via LIDAR or inplace sampling).

A toy example would be the following table, spanning only 5 days for two different regions, A and B:

In [2]:
```
environment_df = pd.DataFrame(dict(
    date=np.repeat(pd.date_range('2021-01-01', '2021-01-05'), 2),
    region=np.tile(['A', 'B'], 5),
    temperature=np.random.normal(20, 5, 10),
    no2=np.random.normal(5, 1, 10),
    fungal_species_1=np.random.normal(1000, 100, 10).astype(int),
    bacterial_species_2=np.random.normal(750, 75, 10).astype(int)))

environment_df.set_index('date')
```

Out[2]:

	region	temperature	no2	fungal_species_1	bacterial_species_2
date					
2021-01-01	A	21.347166	4.983150	1087	828
2021-01-01	B	17.176902	4.450865	965	874
2021-01-02	A	22.338724	4.416838	992	869
2021-01-02	B	16.863509	6.446523	827	780
2021-01-03	A	16.065507	3.814912	894	850
2021-01-03	B	24.239069	4.750215	966	718
2021-01-04	A	11.192028	5.777668	863	798
2021-01-04	B	23.825289	6.599546	1073	796
2021-01-05	A	15.194472	5.062128	1037	851
2021-01-05	B	18.382841	3.336658	1128	832

Healthcare records dataset

The minimal example of a healthcare records dataset that I use would contain, at the individual level, the patient's residence region, and the (vasculitis) onset date recorded.

In [3]:

```python
healthcare_records = pd.DataFrame(dict(
    patient_id=range(1, 16),
    region=np.random.choice(['A', 'B'], 15),
    onset_date=np.random.choice(pd.date_range('2021-01-01', '2021-01-05')
)

healthcare_records.set_index('patient_id')
```

Out[3]:

	region	onset_date
patient_id		
1	A	2021-01-01
2	B	2021-01-01
3	A	2021-01-02
4	B	2021-01-02
5	A	2021-01-01
6	A	2021-01-02
7	B	2021-01-04
8	B	2021-01-04
9	B	2021-01-03
10	A	2021-01-03
11	A	2021-01-04
12	B	2021-01-03
13	B	2021-01-05
14	A	2021-01-03
15	A	2021-01-05

I then go from individual level record to population level records aggregating by date and region, such that the data table I use looks like the following:

In [4]:

```
daily_cases = (healthcare_records
               .groupby(['onset_date', 'region'])
               .size()
               .rename('cases')
               .reset_index()
               .rename(columns={'onset_date': 'date'})
)
daily_cases
```

Out[4]:

	date	region	cases
0	2021-01-01	A	2
1	2021-01-01	B	1
2	2021-01-02	A	2
3	2021-01-02	B	1
4	2021-01-03	A	2
5	2021-01-03	B	2
6	2021-01-04	A	1
7	2021-01-04	B	2
8	2021-01-05	A	1
9	2021-01-05	B	1

Linkage

The final linkage, which leads us to the table on which most of the analyses will be made, is based on merging both the environmental and epidemiological daily incidence counts in a single table based on the `date` and `region` columns, such that:

In [5]:

```
(environment_df
 .merge(daily_cases, on=['date', 'region'], how='left')
 .fillna(0)
 .sort_values(['region', 'date'])
)
```

Out[5]:

	date	region	temperature	no2	fungal_species_1	bacterial_species_2	cases
0	2021-01-01	A	21.347166	4.983150	1087	828	2
2	2021-01-02	A	22.338724	4.416838	992	869	2
4	2021-01-03	A	16.065507	3.814912	894	850	2
6	2021-01-04	A	11.192028	5.777668	863	798	1
8	2021-01-05	A	15.194472	5.062128	1037	851	1
1	2021-01-01	B	17.176902	4.450865	965	874	1
3	2021-01-02	B	16.863509	6.446523	827	780	1
5	2021-01-03	B	24.239069	4.750215	966	718	2
7	2021-01-04	B	23.825289	6.599546	1073	796	2
9	2021-01-05	B	18.382841	3.336658	1128	832	1