Plan Overview

A Data Management Plan created using DMPonline

Title: STRATA-FIT

Creator: Brenda Bley Folly

Principal Investigator: Paco Welsing

Affiliation: UMC Utrecht

Funder: European Commission

Template: Horizon Europe Template

ORCID iD: 0000-0003-2361-2803

Project abstract:

Difficult-to-treat rheumatoid arthritis (D2T RA) is an area of huge unmet medical need with major socio-economic consequences for patients and society. Contributing factors have been identified including co-morbidities, drug-related, biological and behavioral factors. However, identifying these patients with specific underlying and overlapping problems, or patients at risk, is a big challenge in practice.

Currently, treatment decisions are random and not sufficiently patient tailored nor data-driven. Therefore, the STRATA-FIT consortium sets out to develop and validate computational models to identify and stratify D2T RA patients into clinically relevant phenotypes using real world clinical data. We will also measure biomarkers of inflammation to further characterize these phenotypes.Subsequently, we will execute a prospective pilot study to test the effectiveness of a stratified and personalized treatment strategy among others based on our results aided by an online decision support tool.

In parallel we will develop a computational model to identify early RA patients at risk of developing D2T RA. By doing so we cannot only provide better treatment for patients with D2T RA but also work towards its prevention in early RA patients. STRATA-FIT will establish a unique European Learning Healthcare System, using a privacy-proof, state-of-the-art federated learning infrastructure in which patients with, or at risk of D2T RA are identified, stratified and treated in a personalized manner. STRATA-FIT builds on previous work by consortium partners, who initiated and lead the European Task Force developing points to consider for managing D2T RA. It brings together clinical experts, patient research partners and clinical-, biological-, data- and computer-scientists to tackle this major clinical challenge. When successful, STRATA-FIT will lead to more (cost-) effective D2T RA care and will greatly improve the quality of life of D2T RA patients while lowering the burden of D2T RA on Europe’s health care systems and society.

ID: 122183

Start date: 01-05-2023

End date: 04-03-2029

Last modified: 23-06-2026

Grant number / URL: 101080243

Copyright information:

STRATA-FIT

Data Summary

Will you re-use any existing data and what will you re-use it for?

STRATA-FIT will re-use data as collected within Electronic Health records and registries within 5 participating clinical centers. This data will be used to develop and validate clinical prediction models for 1) the identification of D2T RA patients at the moment it arises and 2) the risk of developing D2T RA in early phases of the disease. Furthermore, using this data a stratification model is developed using clustering techniques and prediction for clinical stratification/subgrouping of D2T RA patients to enable better personalised treatment of these patients.

What types and formats of data will the project generate or re-use?

Based on above EHR/registry data, within each center a harmonised and curated, pseudomised dataset will be developed using the FAIR data principles, which can than be 'jointly' analysed using the principle of Federate Learning (using the Personalised Health Train (PHT) concept: https://www.health-ri.nl/initiatives/personal-health-train, as implemented by the project partner MDW (https://www.medicaldataworks.com/ and *https://www.go-fair.org/implementation-networks/overview/personal-health-train/.*

MDW is the initiator of the Vantage6 open source project on "priVAcy preserviNg federaTedleArninG infrastructurE for Secure Insight eXchange ".This is the leading open source implementation of the Personal Health Train principles which MDW markets under the RAILWAY name. See https://distributedlearning.ai/

Once the data across all six centres are harmonised and curated, we will use the PHT concept to connect these data. PHT is designed to enable health care professionals, innovators, and researchers to work with health data from various sources. It enables responsible access to data, while ensuring privacy protection and optimal engagement of data owners. The essence of the PHT concept is that a statistical, research or machine learning question (“train”) travels to a FAIR data source “station” rather than data from various sources having to be transported to a central location for processing. Thus, sensitive data remains where it is. Furthermore, each “station” can accept or reject a “train”, thereby keeping complete data control at the station. Only the answers to the questions such as summary data or model parameters in federated machine learning, are shared by the data provider. Such federated machine learning is often an iterative process requiring multiple communications between “stations” to reach an optimal solution. The infrastructure connecting the “trains” and the “stations” and allowing iterative, secure learning has many legal and technical components and is called the “track” in the PHT metaphor.

MDW has implemented the PHT concept in multiple successful research projects by installing and supporting an open-source software suite called Vantage6, which it co-develops with the Dutch Cancer Registry.

The data (types) that we will/aim to include in these so called FAIR data stations are of the domains:

Demographics : Age (rounded) at diagnosis; gender

Disease Modifying Anti Rheumatic Drugs (DMARD) medication (history): medication type (ATC-code) with start-/*stop-t*ime; reason stopping treatment (ineffectiveness vs adverse effects)

Disease activity: e.g. Disease Activity Score (DAS28), its components and derived indices (i.e. P-

DAS28, swollen- tender difference, current dose for glucocorticoids

I**ndicators for problematic management**: e.g. patients VAS for general health/disease activity / physican VAS / medical procedures (imaging? Local injections?)/ Qol measurements

General ‘time stamp’ of measurements: e.g. diagnosis year; visit/measurement time relative to diagnosis rounded in months; diagnosis is 0, linking measurements that are no more than 2 weeks apart to define ‘observations/data lines’ in database)

Socieconomic status; e.g. *education; income; occupation; family size; relational status*

Laboratory (tox); ALAT; ASAT, creatinine, CRP, ESR, Hb, HBA1c, blood count (leuco’s thrombo’s); Seropositivity for rheumatoid factor and anti-CCP

Comorbidity: Common list of comorbidities as collected between the participating centers. See for list codebook version 2.0 or D3.3.

Lifestyle: smoking; length; weight

Other: *e.g. indicators of: self-efficacy coping, expectation, therapy adherence*

Items in Italic are currently not yet available on data nodes.

Overall a stepwise data extraction is followed so that data is available when analysis are performed. Currently (22-04-2026) a second update of the data node with data on more detailed data on disease activity parameters and on comorbidity has been added to enable the analyses on stratification of the RA and D2T RA population and further predictive analyses.

A central codebook for these variables dat should be made available on the data nodes is kept (see D3.3 for current version d.d. 22-04-2026)

Regarding the prospective clinical pilot study (ACCESS STRATA). The exact design including data collection is detailed in the protocol for this study as approved by the respective Medical Ethical Committees in each country and includes more details on similar domains as mentioned above.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

This data will be used to develop and validate clinical prediction models for 1) the identification of D2T RA patients at the moment it arises and 2) the risk of developing D2T RA in early phases of the disease. Furthermore, using this data a stratification model using clustering techniques and prediction of clinical stratification/subgrouping of D2T RA patients to enable better personalised treatment of these patients.

Within the project the validated prediction models will be integrated into a clinical decision aid for providing personalised care to D2T RA patients. Using this tool (together with clinical treatment recommendations) will be piloted in a prospective clinical study. Based on results a more definite study can be designed and results will likely aid in the developement of treatment recommendations for D2T RA patients and improved outcomes for these patients with a high medical need.

In ACCESS-STRATA: the collected information will provide initial evidence on the feasibility, effectiveness, and cost-effectiveness of the personalised care strategy aided by the developed decsion aid tool.

What is the expected size of the data that you intend to generate or re-use?

Data from about 50,000 RA patients with clinical follow-up data (expected average follow-up > 6 years) will be collected. This concerns harmonised and curated data and is distributed over different

FAIR data stations within the Federated Learning infrastructure as described above.

IN ACCESS-STRATA we will include 175 patients with D2T RA in the observation phase of the study, and 175 D2T RA patients in the study's intervention phase, expectedly, additionally 88 patients need to be included in the intervention phase (total sample size ~ 263 patients). Follow-up will be between 12 months (for patients not participating in both the observation and the intervention phase) and 24 months (for patients participating in both phases of the study). Exact design regarding including data collection is dependent on initial results regarding stratification and will be finalised with clinical expert and patient input, but will include similar domains as described above.

What is the origin/provenance of the data, either generated or re-used?

For the initial phase of the STRATA-FIT it concerns re-used data. For the prospective clinical pilot study (ACCESS-STRATA) the data concerns generated data in a clinical study (with prospective informed concent).

To whom might your data be useful ('data utility'), outside your project?

Data from this real world longitudinal cohort of RA patients may be of use for studying long-term prognosis, treatment effects and development of D2T RA. Relevant questions within the field of Rheumatology. However, as the data concerns sensitive clinical data, the data can not leave their respective policy space. This is also the reason we use the Federated Learning approach within the cohort. As such only analysis results may be shared, after consultation with the consortium and the agreement of individual data-providers.

ACCESS-STRATA: This prospectively collected data will be collected according to the FAIR principles including a provision for making the data available within a repository including timelines and guidance on who and for what the data can be used. This will be described in the protocol and relevant documents for ethical review of this prospective study.

FAIR data

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

The real world cohort will be described extensively and we will provide information on any persistent identifier when applicable and on meta-data in publications and the STRAT-FIT website to make the cohort known to the wider scientific community. We will also provide documentation on the data specifics, formats and ways/criteria of working with this data within a Federated Learning Approach.

For ACCESS-STRATA, we will make metadata available as well as use a persistent identifier.

We will develop a FAIR Implementation Profile (FIP) for above data sets, i.e. a list of declared technology choices intended to implement each of the FAIR Guiding Principles, made as a collective decision by the members of the STRATA-FIT consortium. This FIP will be described in future versions of this DMP.

2.1. Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

We will provide metadata, i.e. we will choose with great specificity and detail which data elements we are going to collect and what they mean (incl. selecting ontologies etc). This will then serve as the input for technology to convert local data into FAIR data. This process is ongoing and the final decision on (meta)data elements and any standards used will be described in future versions of this DMP and the STRATA-FIT codebook and principles used for making data findable will also be described in the FIP as described under 2.1.

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

We will consider the use of search key-words in our metadata. This will be described in the FIP and future versions of this DMP.

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

We aim to offer our metadata in such a way that it can be indexed (and harvested if applicable). Further details of the metadata will be provided in future versions of the DMP and implementation will be described in the FIP (see 2.1).

2.2. Making data accessible - Repository: Will the data be deposited in a trusted repository?

Not for the retrospective/real-world data (as this concerns sensitive data that needs to remain in their respective ‘policy space’).

For the prospective study, ACCESS-STRATA this will be considered, and will be described in the FIP (see 2.1).

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Not yet, will be done when writing the protocol and submission to the ethical committee of the prospective clinical pilot study, ACCESS-STRATA, as well as in the FIP (see 2.1).

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

See above answer

2.2. Making data accessible - Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

For the real-world cohort: no, see 2.2. above.

ACCESS-STRATA: will be considered and described in the study protocol and FIP, see also 2.2. above.

2.2. Making data accessible - Data:

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The real-world data will not be made directly accessible (as this concerns sensitive data that needs to remain in their respective ‘policy space’). Possibly, via the Federated Learning approach as implemented in this project, the pseudo anonymised data may be used for specific aims to be decided on by the STRATA-FIT consortium.

For ACCESS-STRATA: This will be described in the protocol and FIP (see 2.1.).

2.2. Making data accessible - Data:

Will the data be accessible through a free and standardized access protocol?

See above.

2.2. Making data accessible - Data:

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Not for the real-world data, see above.

For ACCES—STRATA, this will be described in the protocol and FIP (see 2.1.).

2.2. Making data accessible - Data:

How will the identity of the person accessing the data be ascertained?

The specifics of processing and managing of making data accessible will be addressed in future versions of the DMP and FIP (see 2.1 and above)

2.2. Making data accessible - Data:

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

As described above, the real-world data will not be made directly accessible as this concerns sensitive data that needs to remain in their respective 'policy space', see above. Possibly, via the Federated Learning approach as implemented in this project, the data may be used for specific aims to be decided on by the STRATA-FIT consortium. For this a data access/FL committee may be needed. However, specifics regarding this will be decided on during the project and will be described in future versions of this DMP and FIP (see 2.1).

For ACCESS-STRATA: This will be described in the protocol and FIP (see 2.1.).

2.2. Making data accessible - Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Metadata will be made openly available. The specifics will be defined in the FIP. See also answers above.

2.2. Making data accessible - Metadata:

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

These specifics will be defined in the FIP and in future versions of the DMP.

2.2. Making data accessible - Metadata:

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

For real-world data, the Federated Learning infrastructure as implemented may be used to analyse the combined data (with permission of data providers). User manuals or similar will be provided for this.

For real-world data and ACCESS-STRATA: details will be defined in FIP and in future versions of the DMP. See also above.

2.3. Making data interoperable:

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

We will use existing standards as much as possible. Details will be defined in FIP and in future versions of the DMP. Also see above.

2.3. Making data interoperable:

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

See answer above.

2.3. Making data interoperable:

Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1] A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

We will make references to the individual data sources for the (FAIR) data stations as defined within the Federated Learning infrastructure of the current project, when possible.

2.4. Increase data re-use:

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

See above. A codebook and Statistical Analysis Plan (SAP) will be made available. Further details will also be made available in FIP and in future versions of the DMP.

2.4. Increase data re-use:

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

No for the real-world data.

For ACCESS-STRATA: we will consider making data freely available in the public domain. Further details will also be made available in FIP and in future versions of the DMP.

2.4. Increase data re-use:

Will the data produced in the project be useable by third parties, in particular after the end of the project?

As much as possible, see answer 2.2.

2.4. Increase data re-use:

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, this will be described throughout the FIP and using references to original data sources as well in describing the real world data e.g. by meta-data and in the protocol of ACCESS-STRATA. With more details in future versions of this DMP.

2.4. Increase data re-use:

Describe all relevant data quality assurance processes.

Data curation and harmonisation steps will be described in detail, among others in FIP and in defining meta-data. Information will be updated in future versions of this DMP.

2.4. Increase data re-use:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Only pseudo anonymised or anonymised data will be used. For the real-world cohort data, a privacy preserving technique for jointly analysing data will be used and in the prospective study informed consent will be asked also describing the reuse of data (FAIR principles). See further 2.1 above

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

We will develop a medical device (decision aid). Although full-CE marking will probably not be performed during the project a Medical Device dossier in line with the Medical Device regulation will be provided for future CE marking.

We will also establish a biobank alongside the prospective pilot study (ACCESS-STRATA) which will be properly documented.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-

use, in line with the FAIR principles.

We will describe the decision aid in sufficient detail to judge its validity and feasibility and details of the biobank including measurements performed and procedures and regulations regarding the use of biobank data will be defined and be described in more detail in future versions of this DMP.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

Not clear at this stage of the project.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Cost for data and documentation/other outputs sharing are covered by the project during the project. A plan will also be devised during the project how/if we can continue the Real-World cohort including the FL infrastructure (including budget needed).

Who will be responsible for data management in your project?

WP1 together with WP2 and WP3 and WP5 for the prospective pilot study (ACCESS-STRATA)

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

A plan for this will be devised during the project including budget needed, see above

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

For the real-world cohort data, local practices for data security are in place in line with local and EU regulations. Source data remains in the Electronic Health records of the participating clinical centres (and as such can be used for data recovery). We will use a federated learning approach to enable

analysis of this sensitive data as made accessible in fair data stations, see above).

Data for ACCESS-STRATA will be collected using the electronic data capture tool Castor (www.Castor.nl) this is a secure cloud-based EDC system including e.g. regulated access to (subsets of) the data, an audit trial and possibilities to mitigate data-entry errors. UMC Utrecht has an approved Data Protection Impact Assessment and a processing agreement with Castor EDC. This system thus enables secure storage/archiving and transfer of sensitive data.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

See above in e.g. 2.2. For the Real-World cohort, these issues will be accounted for in the Federated Learning infrastructure (and e.g. local DPIAs).

For the prospective clinical pilot ACCESS-STRATA, this will be considered within the protocol to be developed.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

For ACCESS-STRATA this will be included.

For the real-world data this is arranged per data provider (using varying informed consent procedures).

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

Not applicable