Food for Thought ...

Guidelines for FAIR Sharing of Preclinical Safety and Off-Target Pharmacology Data

Katharine Briggs[1], Nicolas Bosc[2], Tima Camara[1], Carlos Diaz[3], Phil Drew[4], William C. Drewe[1], Jan Kors[5], Erik van Mulligen[5], Manuel Pastor[6], Francois Pognan[7], Jordi Ramon Quintana[6], Sirarat Sarntivijai[8] and Thomas Steger-Hartmann[9]

[1]Lhasa Limited, Leeds, UK; [2]EMBL-EBI, Wellcome Genome Campus, Cambridge, UK; [3]Synapse Research Management Partners S.L., Barcelona, Spain; [4]PDS Consultants, Leicester, UK; [5]Dept. of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [6]GRIB, Hospital del Mar Institute of Medical Research (IMIM), Dept. of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain; [7]Novartis Pharma AG, Novartis Institutes for Biomedical Research, Basel, Switzerland; [8]ELIXIR Hub, Wellcome Genome Campus, Cambridge, UK; [9]Bayer AG, Research & Development, Pharmaceuticals Investigational Toxicology Building, Berlin, Germany

Abstract

Pre-competitive data sharing can offer the pharmaceutical industry significant benefits in terms of reducing the time and costs involved in getting a new drug to market through more informed testing strategies and knowledge gained by pooling data. If sufficient data is shared and can be co-analyzed, then it can also offer the potential for reduced animal usage and improvements in the *in silico* prediction of toxicological effects. Data sharing benefits can be further enhanced by applying the FAIR Guiding Principles, reducing time spent curating, transforming and aggregating datasets and allowing more time for data mining and analysis. We hope to facilitate data sharing by other organizations and initiatives by describing lessons learned as part of the Enhancing TRANslational SAFEty Assessment through Integrative Knowledge Management (eTRANSAFE) project, an Innovative Medicines Initiative (IMI) partnership which aims to integrate publicly available data sources with proprietary preclinical and clinical data donated by pharmaceutical organizations. Methods to foster trust and overcome non-technical barriers to data sharing such as legal and IPR (intellectual property rights) are described, including the security requirements that pharmaceutical organizations generally expect to be met. We share the consensus achieved among pharmaceutical partners on decision criteria to be included in internal clearance procedures used to decide if data can be shared. We also report on the consensus achieved on specific data fields to be excluded from sharing for sensitive preclinical safety and pharmacology data that could otherwise not be shared.

1 Introduction

Pre-competitive data sharing can offer the pharmaceutical industry significant benefits in terms of reducing the time and costs involved in getting a new drug to market through more informed testing strategies and knowledge gained by pooling data. If sufficient data is shared and can be co-analyzed, then it can also offer the potential for reduced animal usage and improvements in the *in silico* prediction of toxicological effects (Briggs, 2018). Other cross-company initiatives for pre-competitive sharing of pharmaceutical preclinical toxicology data include the IMI eTOX project[1] and BioCelerate[2].

Data sharing benefits can be further enhanced by applying the FAIR Guiding Principles (Wilkinson et al., 2016), which aim to make data more findable, accessible, interoperable and reusable for humans and also for machines. FAIRification of data can offer time savings when reusing data in terms of curating, transforming and aggregating datasets, thereby allowing more time for data mining and analysis (Wise et al., 2019). In addition, application of global, harmonized, comprehensive and open data standards enables interoperability, linkage of data and understanding of data quality, which will support regulatory acceptability of derived evidence based on the data (Cave et al., 2020).

[1] http://www.etoxproject.eu/

[2] https://transceleratebiopharmainc.com/biocelerate/

Received November 18, 2020; Accepted February 16, 2021;
Epub February 25, 2021; © The Authors, 2021.

ALTEX 38(2), 187-197. doi:10.14573/altex.2011181

Correspondence: Katharine Briggs, PhD
Lhasa Limited, Granary Wharf House
2 Canal Wharf, Holbeck, Leeds LS11 5PS, UK
(katharine.briggs@lhasalimited.org)

The Enhancing TRANslational SAFEty Assessment through Integrative Knowledge Management (eTRANSAFE) project[3] is an Innovative Medicines Initiative[4], a public-private collaboration between eight academic research institutes, six small-medium enterprises and twelve pharmaceutical organizations, which aims to integrate different types of data utilized for drug safety assessment in a holistic manner. It combines publicly available data sources with proprietary preclinical and clinical data donated by pharmaceutical organizations. One of the use cases being explored for this data is the generation of virtual control groups, which could be used to reduce the number of animals required for concurrent control groups (Steger-Hartmann et al., 2020).

A key deliverable that supports the sustainability of the project output is the development of policies and guidelines for effective sharing of proprietary data. Beyond developing guidelines that can be practically implemented within the project and used to overcome company-internal resistance to data sharing, the eTRANSAFE project also aims to publicize them in order to gain widespread adoption and reuse by other data sharing initiatives, regulatory bodies and international standard-setting organizations. A general framework for data sharing guidelines which were developed in eTRANSAFE is described in a separate publication (unpublished data Sirarat Sarntivijai et al.). By describing our lessons learned, we hope to facilitate data sharing by other organizations and initiatives and potentially lead to the development of a standard operating procedure (SOP) for FAIR data sharing of preclinical safety and off-target pharmacology data.

2 Findable

The first step in making data FAIR is to make it findable in order to assist discovery and reuse by third-parties. That is to say, the data can be identified unambiguously when looking for it using common search strategies. The FAIR guidelines make reference to globally unique and persistent identifiers, rich metadata, and registration or indexing in a searchable resource, e.g., the internet. What does that mean in terms of making data findable? Imagine you wanted to find an old school friend named John Smith. If you put his name into a search engine, you would get a lot of hits. But if John Smith had joined a searchable resource and, among the metadata, he had added a picture of himself and reference to the school he attended, you can see that he becomes more findable. It would be better still if there was only one John Smith in the world, i.e., if his name was in fact a globally unique and persistent identifier. Persistent identifiers (Philipson, 2019) that can be used for data include URLs, DOIs, ARKs, Handles and ORCiDs. For dynamic datasets, it will in fact be necessary to have a separate global, unique and persistent identifier for each released version (Lamprecht et al., 2020).

What is meant by the term "rich metadata"? Metadata is defined as "data that provides information about other data" or "data about data". Rich metadata is about assigning sufficient descriptive information to help us find and reuse the data. For instance, answering the what, when, where, who, how, which and why questions for the data (Ram and Liu, 2009). Increasing the amount of machine-readable metadata provided for a dataset will also increase the likelihood that a web search will find it. Metadata can usually be shared with other scientists even if access to the data itself needs to be restricted for reasons of sensitivity. It informs others of research that has been done and that the data exist, even if they are too sensitive to be shared beyond the original research team. The FAIR principles recommend that metadata should persist even after the data has been removed, and in this case the metadata should ideally include a statement about when and why the data was removed.

Standards for reporting dataset metadata to aid findability of datasets on the internet include the Schema.org Dataset markup[5] and W3C's Data Catalog Vocabulary format[6], which are used to power Google's dataset search[7]. Bioschemas, which aims to improve findability of life science data, also utilizes Schema.org markup[8]. The UK Data Archive (Corti et al., 2019) has also issued recommendations for metadata to help potential users find datasets and judge whether they are suitable for their research purpose. Other resources that can be used by potential users to locate suitable datasets include Fairsharing.org[9], which provides a registry of databases described according to the BioDBCore database standard. BioDBCore is a community-defined, uniform, generic description of the core attributes of biological databases[10].

In terms of findability, a task force has been set up within eTRANSAFE to look at the sustainability of the project results including long-term access to the proprietary data being shared within the project.

3 Accessible

It is important to understand that making data FAIR does not necessarily mean making data open. Sensitive information can still be protected when implementing the FAIR Guiding Principles, provided that the communication, authentication and authorization protocols utilized adhere to open standards and are clearly defined (Wise et al., 2019). For example, the OpenAPI specification[11] is a standard programming language-agnostic interface description

3 https://etransafe.eu/
4 https://www.imi.europa.eu/
5 https://schema.org/Dataset
6 https://www.w3.org/TR/vocab-dcat/
7 https://developers.google.com/search/docs/data-types/dataset
8 https://bioschemas.org/
9 https://fairsharing.org/
10 https://www.biocuration.org/community/standards-biodbcore/
11 https://www.openapis.org/

for HTTP (hypertext transfer protocol) application programming interfaces (APIs) including REST (representational state transfer) APIs. However, getting agreement on acceptable authentication and authorization protocols can be a challenge when dealing with geographically distributed organizations that have different internal security organization guidelines.

The conditions under which the data can be used should also be clear, both to humans and to computers. Having a standardized and machine-readable way to request access to the datasets as well as standardized descriptions of the data license and terms of use would also be beneficial.

In an ideal world, all data would be shared with the wider scientific community, however, it is recognized that for reasons of sustainability and due to the sensitive nature of the data, it may only be possible for it to be shared with restricted groups made up of trusted partners. In this case, preservation of trust and secure management of information is paramount. Within eTRANSAFE, this has been achieved through the assignment of an independent third party as honest broker, along with a well-defined consortium agreement that all partners have agreed and signed. International collaborative projects and public-private partnerships have developed many different legal agreements for data sharing that differ not only among projects but can often differ within a project among the participating partners. There is currently neither a broadly accepted procedure for safe data sharing nor are contractual templates available.

The honest broker's role is to facilitate data sharing by acting as a trusted neutral partner, hosting the shared data, and controlling data access in accordance with the wishes of all the data owners. The responsibilities associated with this role include collection, curation, and secure management of the shared data. The role may also extend to facilitating discussions around the rules of engagement for the data sharing group and minimum quality criteria that shared data needs to meet. Where information on the data owner is blinded, the honest broker can also act as an intermediary for data access requests. It is important to note, however, that the data remains the intellectual property of the data donor.

Trust is an essential prerequisite for data sharing, and therefore an essential requirement for the honest broker is their independence. A legal framework is necessary to ensure that the honest broker cannot be influenced or commercially dominated by individual partners or third parties in a way that could compromise the data security or confidentiality status of the shared data.

3.1 Security procedures

Appropriate security procedures to prevent unauthorized access and unauthorized changes to the data are vital to maintain the needed levels of trust. These procedures need to be proportionate to the risks involved and agreed with the data owners, encompassing physical security, network security, security of computer systems and files, as well as legal agreements and contracts.

In general, data donors expect the level of security and privacy protection to match or exceed what is implemented in their own IT environment, and the requirements are expected to evolve in line with security best practices. A survey of pharmaceutical partners involved in eTRANSAFE suggested that compliance with the International Organization for Standardization/International Electrotechnical Commission 27001 security standard[12] as well as the European General Data Protection Regulation (GDPR) (EU, 2016) and United States Health Insurance Portability and Accountability Act (HIPAA) (US, 1996) for personal data protection was expected. Other standards data donors mentioned in the survey included the International Standard on Assurance Engagements 3402[13], System and Organization Controls 2 Type 2[14], Organisation for Economic Co-operation and Development 17 Application of GLP (Good Laboratory Practice) Principles to Computerised Systems[15], ISO 27017 on guidelines for information security controls applicable to the provision and use of cloud services[16], Health Information Trust Alliance certification[17], and National Institute of Standards and Technology cybersecurity framework[18].

In recent years, local storage of data has migrated to cloud-based technology, and a key requirement of the eTRANSAFE project was support for cloud hosting, but most pharmaceutical partners indicated they would need to audit the cloud vendor. Other expectations included a disaster recovery plan aligned with legal or regulatory requirements, encryption to protect data at rest, signature and encryption for data in transit, and role-based access controls (RBAC). Single sign on (SSO) authentication along with integration with existing enterprise identity management was required for normal users but strong two-factor authentication for administrator roles. Applications were expected to be locked after a pre-determined number of unsuccessful login attempts, to restrict simultaneous logins from the same user ID, and to automatically log a user out of the system after a predetermined amount of inactivity. A high level of data traceability and auditing was also expected, including, as a minimum, logging of the following security-related events:

- Successful and failed attempts to access systems
- Files and networks accessed
- Changes to system configurations

12 https://www.iso.org/isoiec-27001-information-security.html

13 http://isae3402.com/ISAE3402_overview.html

14 https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/aicpasoc2report.html

15 http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2016)13&doclanguage=en

16 https://www.iso.org/standard/43757.html

17 https://hitrustalliance.net/hitrust-csf/

18 https://www.nist.gov/cyberframework

– Use of system utilities
– Exceptions and other security-related events, such as alarms triggered
– Activation of protection systems, such as intrusion detection systems and anti-malware

Software was expected to be developed in accordance with good security practice including code review, automated testing of security controls, and vulnerability scans of the application. Security updates with good patch management for the platform were to be developed, and security was to be tested regularly, e.g., using penetration testing.

Data that needs to be protected includes personal data, such as that regulated under GDPR (EU, 2016) and commercially sensitive data deemed to be intellectual property (IP). In the case of preclinical and off-target binding data being shared in eTRANSAFE, there is no legal basis for the honest broker to be given access to GDPR-related information, and, therefore, the respective data donor is required to remove this data prior to donation. This can be challenging, as for Good Laboratory Practice (GLP) purposes such information will normally reside within the internal laboratory information management system (LIMS) and Standard for the Exchange of Nonclinical Data (SEND) transport files that are used as a source of these donations. For example, study director (STDIR), principal investigator (PINV), sponsor's monitor (STMON), and contributing scientist (CNTRBSC) are submission values in the CDISC (Clinical Data Interchange Standards Consortium) SEND controlled terminology (SEND-CT)[19].

In terms of clinical data, an assessment is being made of possible formats for data sharing, and preference is being given to aggregated data formats, since this will avoid issues around compliance with GDPR and informed consent. For example, the Periodic Safety Update Report (PSUR) is a pharmacovigilance document provided to regulatory authorities that summarizes cumulative information on risk and benefits and is updated at defined time points to take into account new or emerging safety information[20].

A key requirement expressed by the pharmaceutical partners is the ability to be able to see and query all of their own data including sensitive data fields whilst ensuring that this data remains redacted or obscured for other consortium partners. Here, we have used the term "redacted" to indicate data removed prior to upload to the database such that the data is no longer visible and cannot be retrieved by any user, whereas the term "obscured" is used to indicate data that is removed based on role-based access controls, such that the data is visible only to users with the correct access permissions. The benefit of the second option is that, since the data is still present in the database, it supports the requirement for data donors to easily change the status of donated data, allowing previously sensitive data to be made shareable once it becomes less sensitive, e.g., once a drug goes to market.

3.2 Data classification

Data classification allows the data owner to specify who can access a particular data record. At the start of eTRANSAFE, the data donated to the project was expected to be classified as either non-shareable or fully-shareable with other eTRANSAFE partners. Halfway through the project term, as a large portion of the donated data (approximately 90%) was classified as non-shareable, it was decided to investigate whether an additional data category of partially-shareable data could be introduced where specific fields containing data that donors considered too sensitive to share could be redacted or obscured from data records, allowing the rest of the data record to be shared (Tab. 1). Ten of the twelve data donors in the consortium expressed willingness to use the new partially-shareable data category. The value of partially-shareable data can be illustrated by a recent initiative of eTRANSAFE to collect control animal data to explore the possibility of replacing control group animal data sets from a control group repository. Since the control group animal data is not connected to any structural or pharmacological information, there is no intellectual property issue around it, and thus there is a high willingness to share these data for the purpose of the 3Rs (replacement, reduction and refinement of animal experiments) (Steger-Hartmann et al., 2020).

The right balance would be needed in terms of keeping data unredacted so it is available for the project use cases whilst also maximizing the amount of data made shareable by allowing blinding of at least some parts of it. A survey was conducted to determine if a consensus could be reached on the sensitive data that would need to be redacted or obscured. Nine of the twelve pharmaceutical organizations involved agreed to a shortlist of six key criteria for redaction:

– Chemical structure
– Chemical code, internal compound code, name or reference
– Pharmacological target
– Indication(s)
– Off-target *in vitro* panel
– Company name or identifier

Although removing the above information will reduce the number of potential use cases for the data, it does allow sharing of sensitive data and opens up the possibility for the setting up of one-to-one agreements for data access. For clarity, it is recommended that sensitive data are replaced with the term "redacted", so it is clear data have been removed rather than being missing for some other reason.

One partner expressed concern over the value of the data without a structure or descriptors such as logarithm of the partition coefficient (LogP) or volume of distribution, which could help understand the tissue concentration and toxicity finding. However, another partner noted that physicochemical properties would also need to be redacted, as these could be used to derive the chemical structure.

19 https://www.cdisc.org/standards/terminology/controlled-terminology

20 https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-good-pharmacovigilance-practices-module-vii-periodic-safety-update-report_en.pdf

Tab. 1: Data classifications used in eTRANSAFE and who is granted access to this data

Status of eTRANSAFE data	Who can see the data			
	Data donor	Honest broker	Other eTRANSAFE participants	Public, i.e., not restricted to eTRANSAFE use cases
Non-shareable data Data is only accessible to the donor and the honest broker.	Yes	Yes	No	No
Partially-shareable data Data can be shared within the consortium; however, some data will have been redacted/obscured. Data cannot be shared outside of the consortium.	Yes	Yes	Yes	No
Fully-shareable data Data can be shared within the consortium. Data cannot be shared outside of the consortium.	Yes	Yes	Yes	No
Public data Data use is not restricted.	Yes	Yes	Yes	Yes

One partner expressed willingness to share their company name in order to facilitate the setting up of one-to-one data sharing agreements. Another partner requested that site-specific study identifiers be added to the list of identifiers, which could include test subject identifiers and site-specific animal strains. However, redaction of these fields would significantly impact on usability of the data, since it would then not be possible to aggregate the data on a per subject or per study basis. An alternative solution would be for the data donor to anonymize this information. This partner also flagged that some laboratory tests can be test substance-specific. Other candidates for anonymization within the SEND standard are the long (PCTEST & PPCAT) and short (PCTESTCD) names for the analyte in the pharmacokinetics concentrations (PC) and pharmacokinetics parameters (PP) domains. If the raw data could be replaced by harmless terms such as "parent" or "metabolite", this could substantially increase the value of the data.

3.3 Clearance procedures

Before pharmaceutical organizations can release proprietary data to third parties, the request to share data needs to go through an internal clearance process. In the case of clinical data, most companies have set up internal operating procedures for how such requests should be handled, utilizing a central assessment team or review board. The key aspects which need to be assessed are data protection according to the General Data Protection Regulation, whether existing informed consent allows data re-use and sharing, as well as the commercial sensitivity of the efficacy results obtained. Decisions should be guided by the joint EFPIA and PhRMA principles for responsible clinical trial data sharing[21].

Bayer AG have shared information on their internal clearance procedure for clinical data as an example. The review board is coordinated by a data sharing coordinator and consists of the following core functions; therapeutic area, clinical development, clinical statistics, law & patents, medical affairs and project management. Requests need to be submitted in writing and include the following information:

- Short project description including rationale and objectives for data sharing
- Identification of the honest broker who will host the data
- Data sharing agreement (if there is any for a specific project)
- Identification of the desired data sets and endpoints (laboratory data, adverse events, efficacy results and/or others)
- Identification of the intended aggregation level

By contrast, the internal clearance procedures to release non-human data are often managed on a case-by-case basis, are time consuming, and are assigned a low priority because they are not part of daily business operations. It was recognized during the IMI eTOX project that such processes were a barrier to data sharing and would benefit from simplification, for instance, by having an SOP for obtaining authorization that identifies the steps needed, including clear roles and responsibilities as to who owns the data, who can authorize data sharing and on what basis. Centralizing the process as much as possible could also speed up such decisions and would ensure a harmonized decision-making process.

21 https://www.efpia.eu/media/25666/principles-for-responsible-clinical-trial-data-sharing.pdf

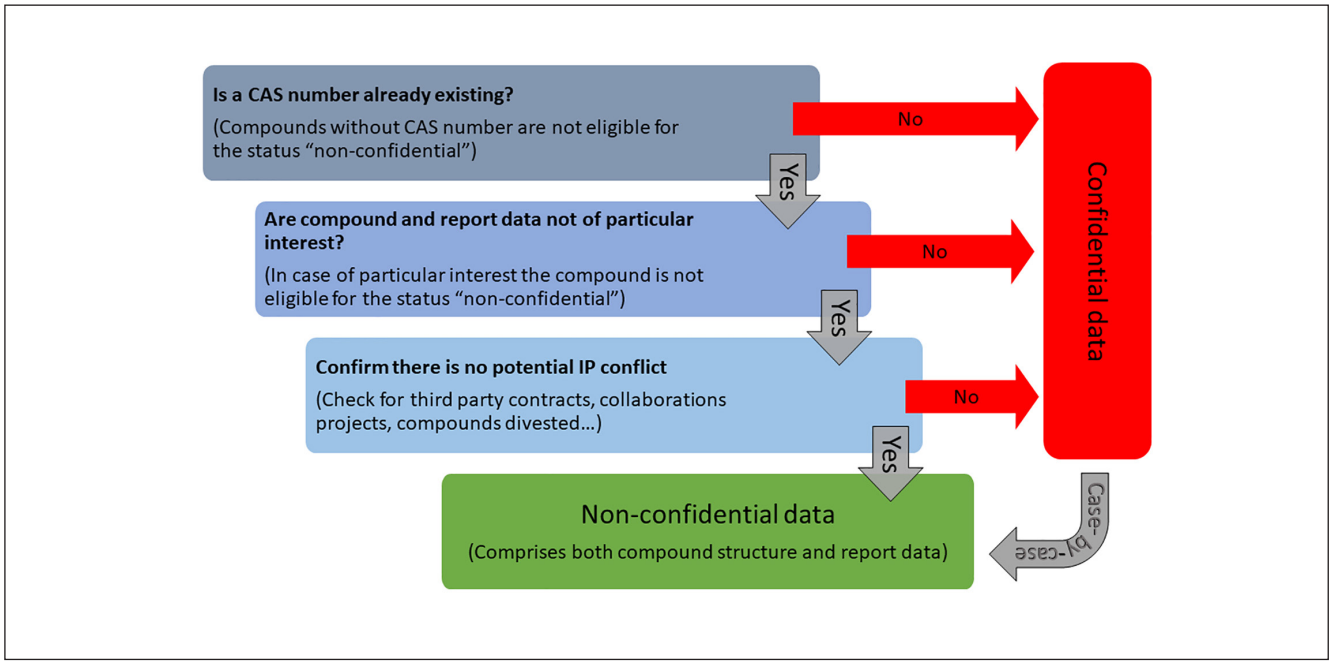

Fig. 1: Clearance procedure for preclinical data implemented at Bayer AG

A survey was carried out to assess how such requests are currently managed by the different pharmaceutical organizations participating in eTRANSAFE and to assess if a consensus could be reached. The clearance procedure implemented by Bayer AG during the IMI eTOX project was put forward as a benchmark (Fig. 1).

Of the twelve pharmaceutical organizations involved in eTRANSAFE, nine provided information on their internal clearance procedures. Of the eight partners in eTRANSAFE who were also partners in the IMI eTOX project, five had taken steps to set up SOPs, one was in the process of implementing such procedures, and two did not provide a response. Of the remaining four eTRANSAFE partners who were not involved in IMI eTOX, two had set up SOPs, one did not respond, and one is planning to redact data they consider to be too sensitive prior to donation. From the survey results, the criteria considered as part of the decision-making process include:

Step 1: Whether the data are already publicly available, i.e., data that is either protected by patents or represents prior knowledge; indicated by five partners. Here, one partner used the presence of a CAS (Chemical Abstracts Service) number determined by a central library service as a proxy for data being in the public domain, one partner indicated that all current or past marketed drugs were shareable, and another took into account whether the data had already been shared with other consortia.

Step 2: Whether the compound and study data are still of strategic interest, e.g., the drug is still in development or is being repurposed; indicated by five partners. A key difference here was who was responsible for this decision; roles that were mentioned include head of research, head of development, head of medicinal chemistry, compound development team leader, safety representative, project leader, life cycle manager, patent attorney and head of patent/IP.

Step 3: Whether there is a potential IP or legal conflict; indicated by six partners; two specified this was their primary consideration or first step in the process. Potential conflicts can occur where the data owner does not have the exclusive rights, e.g., joint ownership due to collaborations or divestments or where the drug is under litigation or patent dispute. However, joint ownership does not preclude data sharing but would require the consent of all owners.

Step 4: Changes to the classification; mentioned by three partners, the expectation being that the data owner can change the status and fully share the data at a later time point, i.e., where the data owner has come to the conclusion that the sensitivity has become less critical. Indeed, it is recommended that data classification is reviewed periodically in order to facilitate this step.

Procedures for granting third party access have not yet been elaborated in eTRANSAFE but are likely to build on the procedure set up for the IMI eTOX project. In both the eTOX and eTRANSAFE projects, the data donor retains ownership of their own IP. Therefore, to gain access, third parties needed to negotiate with each data donor whose data was of interest. In order to simplify this process and to ensure such requests were handled consistently, an internal review group was set up during the eTOX project. This group reviewed third party access requests submitted using a standardized request form and then issued a recommendation based on the perceived benefits to the project.

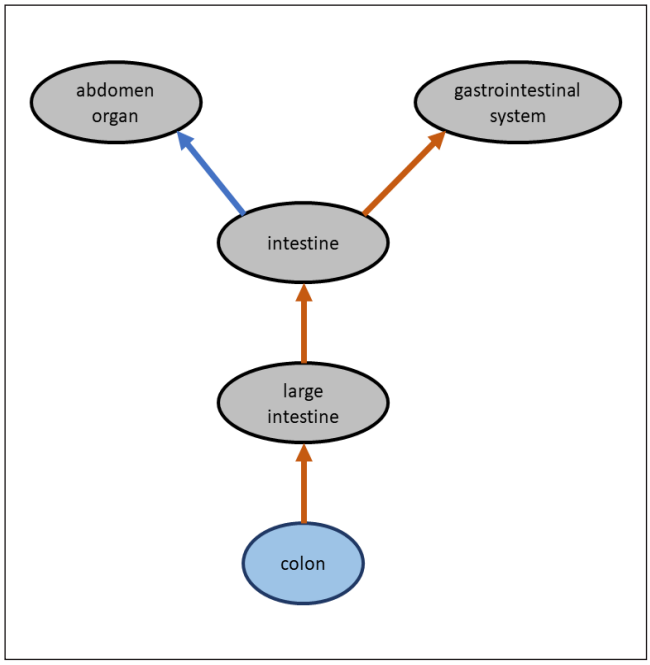

Fig. 2: Relationship between colon and gastrointestinal system in the eTOX anatomy ontology

4 Interoperable

To be interoperable, data should be machine-readable and use terminologies, vocabularies, or ontologies that are open, globally applicable, and commonly used in the field. In addition, using platform-independent data standards and formats can help to protect against obsolescence of hardware and software environments. However, the choice of an appropriate standard is not always clear. For instance, there are several standards and formats for capturing clinical data, e.g., CDISC Standards, Health Level Seven[22], Periodic Safety Update Report (PSUR), Development Safety Update Report[23] and Investigator's Brochure[24].

Even in the case of preclinical data, where data formatting has coalesced around the CDISC SEND standard[25], there are still challenges because this standard is evolving. In order to be interoperable, data formats would need to be adapted to newer versions of the SEND standard, for instance SENDIG DART v1.1[26], whilst still maintaining back compatibility with the existing data.

For other data types, there may be no clear standard, and one will need to be defined. For instance, within eTRANSAFE, we have developed the SR (Study Report) domain (Drew et al., 2019). This domain was added in order to capture the expert interpretation of findings found in study reports but missing from SEND, which is aimed at capturing the raw individual animal data. This expert interpretation as to whether findings are considered treatment-related or not, including effect levels such as a lowest observed adverse effect level (LOAEL) or no observed adverse effect level (NOAEL), is considered essential to allow the data to be used more widely, particularly by non-pathologists/toxicologists, e.g., for building and validating *in silico* models. Although it is possible to identify abnormal findings by deriving reference ranges and background incidence rates from the raw control data (Pinches et al., 2019), toxicology/pathology expertise is needed in order to assess if the changes observed are adverse or non-adverse; for instance, if the change is adaptive or transient (Lewis et al., 2002).

In the case of the proprietary off-target pharmacology data being shared in the eTRANSAFE project, the plan is to process this data using the same standardization protocols used for data included in the publicly available ChEMBL database so that the data can be utilized in the same manner.

Use of controlled vocabularies and ontologies is recommended in order to avoid differences in spelling and terminology and to enable qualitative findings to be searched in a consistent manner across the different data sources. Ontologies offer additional benefits over vocabularies in that the relationships – synonyms, meronyms/homonyms and hyponyms/hypernyms – between terms can also be captured. This can help where findings are reported at different levels of granularity, e.g., gastrointestinal system versus colon (Fig. 2).

Several controlled vocabularies are applicable in the clinical domain, e.g., MedDRA[27], SNOMED CT[28], LOINC[29], MeSH[30] and RxNorm[31], some of which are proprietary in nature. In the preclinical domain, the CDISC SEND Controlled Terminology (SEND-CT) predominates. This standard is maintained by the National Cancer Institute Enterprise Vocabulary Service[32] with the assistance of the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) Global Editorial Steering Committee (GESC) (Keenan et al., 2015). SEND-CT is also not a static standard, and staying up-to-date with the current version whilst maintaining back compatibility will be a challenge. An example of the difficulties faced is the

22 https://www.hl7.org/about/index.cfm?ref=nav
23 https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use_en-26.pdf
24 https://ichgcp.net/7-investigators-brochure
25 https://www.cdisc.org/standards/foundational/send/sendig-v31
26 https://www.cdisc.org/standards/foundational/send/sendig-dart-v11
27 https://www.meddra.org/
28 http://www.snomed.org/snomed-ct/why-snomed-ct
29 https://loinc.org/
30 https://www.nlm.nih.gov/mesh/meshhome.html
31 https://www.nlm.nih.gov/research/umls/rxnorm/index.html
32 https://www.cancer.gov/research/resources/terminology/cdisc

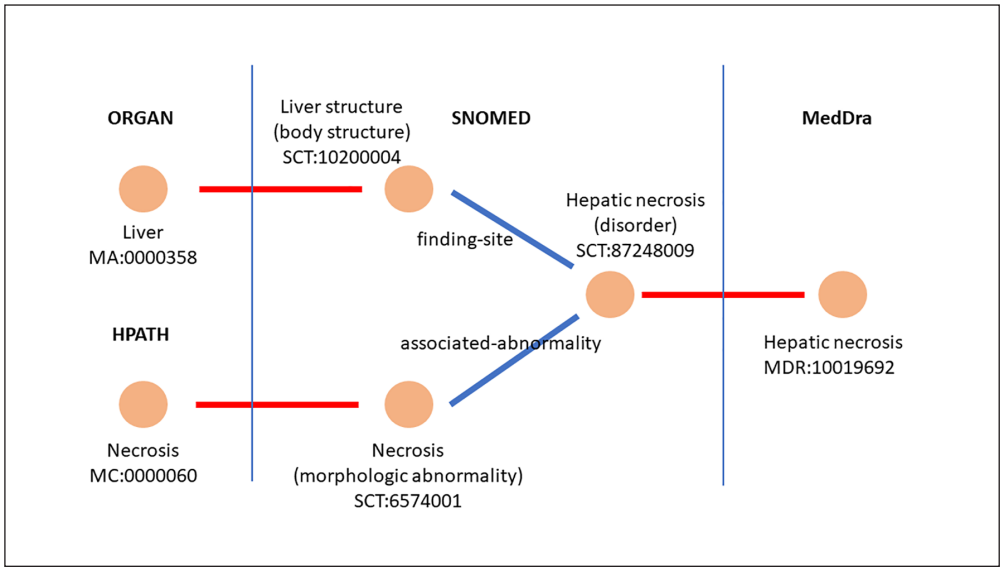

Fig. 3: Example mapping between preclinical and clinical terms using SNOMED CT as an intermediate ontology

change in vocabulary used to express severity of histopathology findings. Previously, SEND-CT used terms such as mild, moderate and severe based on a standardized scale. However, this does not reflect the heterogeneity of the real data, and the latest version of the SEND-CT has moved to expressing this in the form "2 of 4", "3 of 5", etc.

Several ontologies and vocabularies were developed and published as part of the eTOX project, including an ontology of histopathological morphologies[33]. The benefits of including these ontologies and vocabularies in eTRANSAFE is that they were supplemented with a vast number of synonyms encountered during the extraction of the preclinical study reports, whereas the SEND-CT includes only a limited number of synonyms. The histopathology ontology was created, reviewed and updated by pathologists from multiple organizations and was aligned and cross referenced to INHAND terms published during the lifetime of the eTOX project.

With the diversity of endpoints pooled in eTOX, it proved difficult to have a one-size-fits-all vocabulary. Based on this learning, it is not optimal to have an ontology of everything to describe data in the scope of eTRANSAFE. The Semantic Services module in eTRANSAFE will act as a "Rosetta Stone", translating queries into the preferred terms utilized by the different data sources. Several ontologies and controlled vocabularies will be incorporated into the system including the SEND-CT and eTOX histopathology ontology. SNOMED CT is being used as an intermediate ontology to allow mapping between preclinical and clinical terms (Fig. 3).

In addition, OntoBrowser (Ravagli et al., 2017), an open-source collaborative tool for curation of ontologies, will be used to allow for addition of new synonyms and preferred terms which cannot be mapped to an existing ontology. It is already clear that a significant number of the terms utilized within the data donors' LIMS differ from those encountered in the study reports extracted within eTOX.

When applied to different types of data sources, interoperability needs to be extended to encompass the various entities and relationships being captured. In the clinical domain, the Observational Medical Outcomes Partnership (OMOP) common data model allows data from disparate clinical data sources to be queried simultaneously[34] by transforming the data into a common format. Within eTRANSAFE, the different preclinical and clinical data sources are intended to be accessed via primitive adaptors (PAs), an API (Fig. 4) which provides indexes to the data based on abstract data classes that are common to both preclinical and clinical data sources and relevant to the project use cases. The data classes allow data from heterogenous data sources to be queried in the same way, unifying the query search parameters for all data sources connected to ToxHub and identifying the types of data available in the system. It also allows for the results of those queries to be aggregated together without the need for data curating or joining. New data sources can be integrated by adding a primitive adapter for that data source and indexing the data using the common data classes. Currently the primary data classes in use are:

– COMPOUND data class, e.g., RxNorm, InChI
– STUDY data class, e.g., species, route of administration
– FINDING data class, e.g., organ, adverse event

A key requirement for eTRANSAFE is the ability to perform compound structure searching, e.g., exact, similarity or substructure searches. Since chemical structures can be drawn and represented in several different ways, it will be necessary to identify

[33] https://github.com/Novartis/hpath

[34] https://www.ohdsi.org/data-standardization/the-common-data-model/

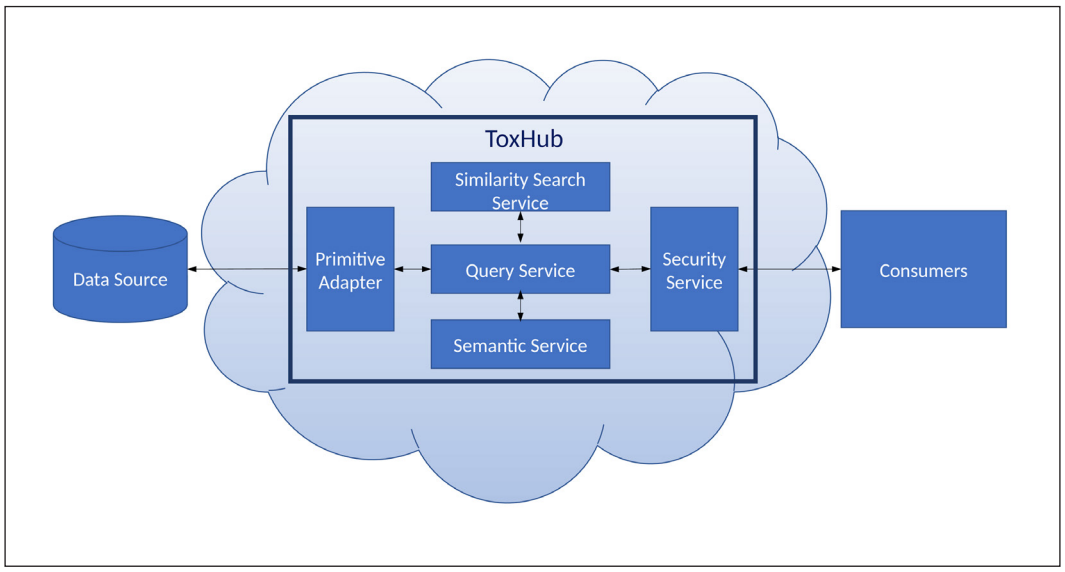

Fig 4: Logical architecture of the ToxHub (iteration 1)

a standardized machine-readable representation[35] to ensure consistency across the different data sources, not only for data entry but for data retrieval, i.e., to decide if drugs from different data sources are identical. For small molecules, several formats are available, e.g., SMILES, Mol files, InChI and InChIKey (Warr, 2011). Possible formats for representation of biologicals include PDB[36], FASTA[37], HELM (Zhang et al., 2012) and SCSR (Chen et al., 2011).

In order to meet the translational aims of the project, it will be necessary to determine the overlap in terms of the compounds associated with preclinical data with those associated with clinical data. UniChem (Chambers et al., 2013) is a freely available system for cross-referencing the chemical structure identifiers used in different databases via InChIKeys. However, it is more common for drugs to be identified by their generic or brand names within clinical data sources. A process to identify the associated structure has been elaborated using the comprehensive list of preferred names available in the US FDA (Food and Drug Administration) Substance Registration System (SRS) list of Unique Ingredient Identifiers (UNII)[38]. These names can then be mapped to chemical structure identifiers used in the ChEMBL database[39]. This process has allowed the inclusion of DailyMed[40] identifiers to UniChem.

Methods for standardizing numerical values to either conventional units or the international system of units (SI) will also be required if quantitative findings are to be queried in a consistent manner. For some units, additional information may be needed in order to perform the conversion, e.g., to convert mass to moles. The first step, however, will be to standardize the units themselves as these can be represented in many ways, for instance, micromoles per litre, umol/l, µmol/l, µmol.l-1, µmol/L, etc.

5 Re-usable

Re-usability of data is one of the key aims of the FAIR Guiding Principles, as it allows data to be repurposed for new user communities, for new needs, and for new applications. Data in this sense can become more valuable to more people across large organizations, whether open-source communities or private organizations. Provisioning long-term access to data requires resources and funding to continue beyond the initial investment. For a sustainable funding model, the number of parties interested in the data and the value they place on it has to be equal to or greater than the costs of maintaining it, including updates, support and training.

A distinction should be made between static datasets and dynamic datasets, which are continually updated. However, both will require long-term preservation for reproducibility. Static datasets will therefore still require investment to manage replacement of hardware as well as updates and security patches for software, such as operating systems and internet browsers in addition to keeping up-to-date with security best practices. Data generation, data processing, and data storage devices and methods have all changed rapidly in the past decade with advances in the computational domain. In the case of clinical data, there is the

35 https://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf
36 https://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/index.html
37 https://www.bioinformatics.nl/tools/crab_fasta.html
38 https://fdasis.nlm.nih.gov/srs/jsp/srs/uniiListDownload.jsp
39 https://www.ebi.ac.uk/chembl/
40 https://dailymed.nlm.nih.gov/dailymed

added complication of needing to adhere to local data protection requirements, which can also change over time.

To be re-usable, data should be sufficiently well-described with metadata and provenance information so that the data sources can be linked or integrated with other data sources and enable proper citation. At the data dictionary level, metadata should include an explanation of database records, table and field names, data types, code definitions, and classification schemes, including how missing values should be reported. SEND-IG provides a good example of the amount of detail required for a data dictionary.

Metadata to track changes and different versions of the data is also crucial to ensure scientific reproducibility. Semantic versioning consisting of three numbers, representing the current major release, the current minor release of the current major release, and the current patch release of the current minor release (i.e., major.minor.patch), can be used to track different versions of database software.

6 Conclusion

For organizations and initiatives interested in data sharing, the FAIR Guiding Principles can offer additional benefits by reducing time spent curating, transforming and aggregating datasets, thereby allowing more time for data mining and analysis. The principles also allow data to be repurposed for new user communities, for new needs, and for new applications.

In terms of findability, thought should be given to how the data are disseminated to potential users and whether the database should be included in a registry such as Fairsharing.org[9], which could enhance its findability.

Regarding accessibility, the licensing model(s) under which the data will be released will need to be elaborated along with the terms and conditions of use. For proprietary data, each data donor will need to separately agree to the inclusion of their data in the sustainable version of the database. It is clear from the results of the security survey that a high level of security and privacy protection will be required to obtain this agreement, e.g., compliance with ISO27001 22, GDPR (EU, 2016) and HIPAA (US, 1996), and will need to evolve in line with security best practices.

Company-internal clearance procedures were identified as a possible blocker to data sharing due to the absence of SOPs for obtaining authorization. A survey conducted on this topic highlighted that there are significant differences in the decision-making process among the pharmaceutical organizations that took part. However, there was a greater consensus on the criteria used as a basis for these decisions:

- Whether the data are already publicly available, i.e., data that is either protected by patents or represents prior knowledge
- Whether the compound and study data are still of strategic interest, e.g., the drug is still in development or is being repurposed
- Whether there is a potential IP or legal conflict, e.g., joint ownership due to collaborations or divestments

It would benefit the scientific community if these commonly agreed decision criteria could be codified into official guidelines similar to the joint EFPIA and PhRMA principles for responsible clinical trial data sharing.

The new, partially-shareable data category will potentially allow more data to be made accessible, although with certain fields containing sensitive data either redacted or obscured. Here, again, there was substantial agreement on the criteria that were considered too sensitive by data donors and require redaction:

- Chemical structure, chemical code, internal compound code, name or reference
- Pharmacological target, indication(s) and off-target *in vitro* panel
- Company name or identifier

Requirements for interoperability will be met by adhering to existing well-defined standards and vocabularies to ensure interoperability. One of the main challenges here is that the applicable standards and vocabularies are still evolving.

The tools being developed within eTRANSAFE to allow preclinical and clinical data sources to be integrated, namely, the Semantic Services module to link terms, the PAs to link data classes, and UniChem to link chemical structure identifiers, could also be used by other initiatives.

Requirements for reusability can be met by having adequate metadata but will also depend to some extent on whether the number of parties interested in the data and the value they place on it is equal to or greater than the costs of maintaining it, including updates, support and training.

It is hoped that the SR Domain to capture expert interpretation of findings will be adopted and maintained by CDISC as an extension to the SEND standard, since it would have benefits for reusability by allowing the data to be used by non-pathologists/toxicologists, e.g., for building and validating *in silico* models.

The data sharing guidelines developed within eTRANSAFE provide practical solutions for effective sharing of proprietary data, which can help overcome potential company-internal resistance to data sharing. We envisage that these guidelines could be reused by other proprietary data sharing initiatives and hope that they will obtain widespread adoption and recognition from regulatory bodies and international standard-setting organizations. The key recommendations of the eTRANSAFE data sharing guidelines are:

- A tiered data classification scheme to assess the intellectual property aspect and the resulting level of confidentiality in order to allow data to be protected as thoroughly as necessary but also to be shared as widely as practical.
- A structured company-internal clearance procedure in which decisions are centralized at each step in order to speed up and harmonize data sharing requests.
- Appropriate security procedures to prevent unauthorized access and unauthorized changes to the shared data including use of an honest broker where applicable.
- Standardization using open and globally applicable standards to facilitate data interoperability.
- Quality assurance and quality control procedures to ensure that the data are as error-free as is practical, including defining minimum quality standards for the data.

– Capturing information on metadata, traceability and provenance to facilitate reproducibility of computational workflows.
– Sustainability plans to protect against obsolescence of hardware and software environments including provision for access by third parties.

References

Briggs, K. (2018). Is preclinical data sharing the new norm. *Drug Discov Today 23*, 499-502. doi:10.1016/j.drudis.2016.05.003

Cave, A., Brun, N. C., Sweeney, F. et al. (2020). Big data – How to realize the promise. *Clin Pharmacol Ther 107*, 753-761. doi:10.1002/cpt.1736

Chambers, J., Davies, M., Gaulton, A. et al. (2013). UniChem: A unified chemical structure cross-referencing and identifier tracking system. *J Cheminform 5*, 3. doi:10.1186/1758-2946-5-3

Chen, W. L., Leland, B. A., Durant, J. L. et al. (2011). Self-contained sequence representation: Bridging the gap between bioinformatics and cheminformatics. *J Chem Inf Model 51*, 2186-2208. doi:10.1021/ci2001988

Corti, L., Van den Eynden, V., Bishop, L. et al. (eds.) (2019). *Managing and Sharing Data: A Guide to Good Practice*. 2nd edition. Essex: UK Data Archive. https://www.ukdataservice.ac.uk/manage-data/handbook.aspx

Drew, P., Thomas, R. and Capella-Gutierrez, S. (2019). PP07: Consolidating study outcomes in a standardised, SEND-compatible structure. FDA/PHUSE US Computational Science Symposium, June 9-11 2019. Silver Spring, Maryland. https://www.lexjansen.com/css-us/2019/PP07.pdf

EU – European Union (2016). Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. *OJ L119*, 1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Keenan, C. M., Baker, J., Bradley, A. et al. (2015). International harmonization of nomenclature and diagnostic criteria (INHAND): Progress to date and future plans. *Toxicol Pathol 43*, 730-732. doi:10.1177/0192623314560031

Lamprecht, A. L., Garcia, L., Kuzak, M. et al. (2020). Towards FAIR principles for research software. *Data Sci 3*, 37-59. doi:10.3233/DS-190026

Lewis, R. W., Billington, R., Debryune, E. et al. (2002). Recognition of adverse and nonadverse effects in toxicity studies. *Toxicol Pathol 30*, 66-74. doi:10.1080/01926230252824725

Philipson, J. (2019). Identifying PIDs playing FAIR. *Data Sci 2*, 229-244. doi:10.3233/DS-190024

Pinches, M. D., Thomas, R., Porter, R. et al. (2019). Curation and analysis of clinical pathology parameters and histopathologic findings from eTOXsys, a large database project (eTOX) for toxicologic studies. *Regul Toxicol Pharmcol 107*, 104396. doi:10.1016/j.yrtph.2019.05.021

Ram, S. and Liu, J. (2009). A new perspective on semantics of data provenance. Proceedings of the First International Conference on Semantic Web in Provenance Management. *CEUR Workshop Proceedings 526*, 35-40. Washington DC: CEUR-WS.org. http://ceur-ws.org/Vol-526/InvitedPaper_1.pdf

Ravagli, C., Pognan, F. and Marc, P. (2017). OntoBrowser: A collaborative tool for curation of ontologies by subject matter experts. *Bioinformatics 33*, 148-149. doi:10.1093/bioinformatics/btw579

Steger-Hartmann, T., Kreuchwig, A., Vaas, L. et al. (2020). Introducing the concept of virtual control groups into preclinical toxicology testing. *ALTEX 37*, 343-349. doi:10.14573/altex.2001311

US – United States (1996). Health Insurance Portability and Accountability Act of 1996. Public Law 104-191. 110 Stat. 1936. https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf

Warr, W. A. (2011). Representation of chemical structures. *Wiley Interdiscip Rev Comput Mol Sci 1*, 557-579. doi:10.1002/wcms.36

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data 3*, 160018. doi:10.1038/sdata.2016.18

Wise, J., de Barron, A. G., Splendiani, A. et al. (2019). Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discov Today 4*, 933-938. doi:10.1016/j.drudis.2019.01.008

Zhang, T., Li, H., Xi, H. et al. (2012). HELM: A hierarchical notation language for complex biomolecule structure representation. *J Chem Inf Model 52*, 2796-2806. doi:10.1021/ci3001925

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

The authors wish to acknowledge the input of members of the eTRANSAFE Consortium to the development of the general framework for data sharing guidelines and completion of the survey responses reported here. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365 ("eTRANSAFE"). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.