IMI2 GA853989 - ERA4TB
European Regimen Accelerator for Tuberculosis

WP1 –Data and Pipeline Management

D1.3 Instantiation of EU-based Drug Development Information Management (DDIM) system

Lead contributor	Jesús Carretero Pérez (1 – Universidad Carlos III de Madrid)
Other contributors	Víctor J. Sosa (1 – Universidad Carlos III de Madrid) Javier García Blas (1 – Universidad Carlos III de Madrid)

Due date	30/06/2020
Delivery date	07/07/20
Deliverable type	R
Dissemination level	PU

Document History

Version	Date	Authors	Description of Changes
0.1	27-Feb-2020	Víctor J.Sosa Jesús Carretero	N/A – Initial draft document
0.2	27-Mar-2020	Víctor J.Sosa	a) Changes in the DDIM architecture, the structure of layers to access subsystems was modified. b) This version of the document is focused on the definition of the access layer using JWT.
0.3	10-Jun-2020	Víctor J.Sosa	a) The document was extended to describe the complete architecture of DDIM, which is not mainly focused only on the DDIM access management, but the different DDIM layers and dataflow; b) Details on user registration are also included; c) New data DDIM-providers interaction using sFTP is included.
0.4	11-Jun-2020	Jesús Carretero	Formatting and review of contents. Quality check.
0.5	23-Jun-2020	Víctor J.Sosa	Final review and inputs by Work package 1 members

Table of contents

Abstract

This document constitutes an initial report to provide an early status on the deployment of the Drug Development Information Management (DDIM) system being implemented by Work Package 1 in the context of the European Regime Accelerator for Tuberculosis (ERA4TB) project. This version of the document is focused on providing a preliminary general overview of the DDIM architecture, describing every DDIM layer and its basic data flow. The DDIM is a single-entry point via a web portal (Common User Interface) that will allow ERA4TB Consortium members to access information on the drug development process and the compounds under investigation. In addition, the data within the DDIM will support a number of downstream activities (e.g. data modelling) in order to meet the primary objectives of the ERA4TB project.

The DDIM wraps the access to separate subsystems that support the management of information during the drug development process. These subsystems include: (1) a Pipeline Management Tool; (2) Preclinical and Clinical Data Capture Systems; and (3) an Image Storage Repository. Additionally, the DDIM will be able to connect to a Data Analysis system to enable the efficient application of machine learning methods to data extracted from the different subsystems. The data within the DDIM will subsequently be made available within a Data Archive platform which will be made available to approved researchers both during and beyond the existence of the ERA4TB project.

1. Introduction

The European Regime Accelerator for Tuberculosis (ERA4TB) project is a research collaboration funded under the Innovative Medicines Initiative Joint Undertaking (IMI JU) within the framework of the wider Antimicrobial Resistance Accelerator programme. The objective of this project is to accelerate the development of new treatment regimens for tuberculosis. ERA4TB will address this aim by bringing drug candidates and combinations ready to Phase II clinical evaluation.

The primary objective is to create a European open platform to accelerate the development of new combination regimens for the treatment of TB. The project will start from a collection of anti-TB compounds resulting from different EFPIA drug discovery activities which are in varying stages of development, and progress the most promising compounds through an ERA4TB 'pipeline' comprised of a variety of in vitro assays, in vivo studies and clinical (Phase I) trials as appropriate for each compound asset.

To cope with all the steps of drug development and manage all data related to each step, a Drug Development Information Management (DDIM) system is being developed to support the entirety of the pipeline workflow and the associated data flow will be designed and deployed by C-PATH, grit42 and UC3M. The DDIM consists of the following major components:

1) A web-based portal (Common-User Interface) will be developed by UC3M to allow Consortium members to access the clinical and preclinical databases and the imaging repository from a single interface. The interface will provide a centralized access control system to ensure secure and efficient management of the data according to distinct user permissions as needed. The interface will allow users within the project to access the available tools for the interrogation, analysis and download of data in the preclinical and clinical data platforms and the imaging repository.

2) A clinical data management platform will be utilised using the grit42 system to manage the clinical data generated by the ERA4TB project and provide tools to interrogate and analyse the data. The existing TB clinical data at C-PATH (i.e. TB-PACTS) will be delivered into the clinical data platform and integrated with clinical data generated from the ERA4TB activities. The inclusion historical data along will further support downstream activities such as data modelling, simulation and machine learning. All data delivered within the platform will be in an agreed dataset structure.

3) A preclinical data management platform using the grit42 system will also be utilised, this instance of grit42 will also possess tools to interrogate and analyse the data. The data resulting from the experiments in preclinical data stages will be collected and stored within the platform using agreed standard data template structures. Similar to the clinical data management platform, C-PATH will be providing pre-existing TB data into the platform to support various downstream activities.

4) An imaging data capture repository (XNAT) will be deployed by UC3M to store and analyse images resulting from partners' experiments in preclinical data stages within the ERA4TB project.

5) A Pipeline Management Tool is being developed by UC3M to support the oversight, decision-making and review of drug assets progressing through the drug development pipeline. Section 3 of D1.1 includes a detailed technical description of the Pipeline Management Tool.

2. Architecture

The DDIM system has been designed based on a layered and modular architecture. This type of architecture allows grouping different components of the system (called subsystems) at different levels that can communicate each other through application program interfaces (APIs) or web services. Every layer is represented by a set of modules or subsystems of varying complexity. Details about the internals of the layers and modules are abstracted and simplified, facilitating the system use and integration. Figure 1 shows the DDIM architecture.

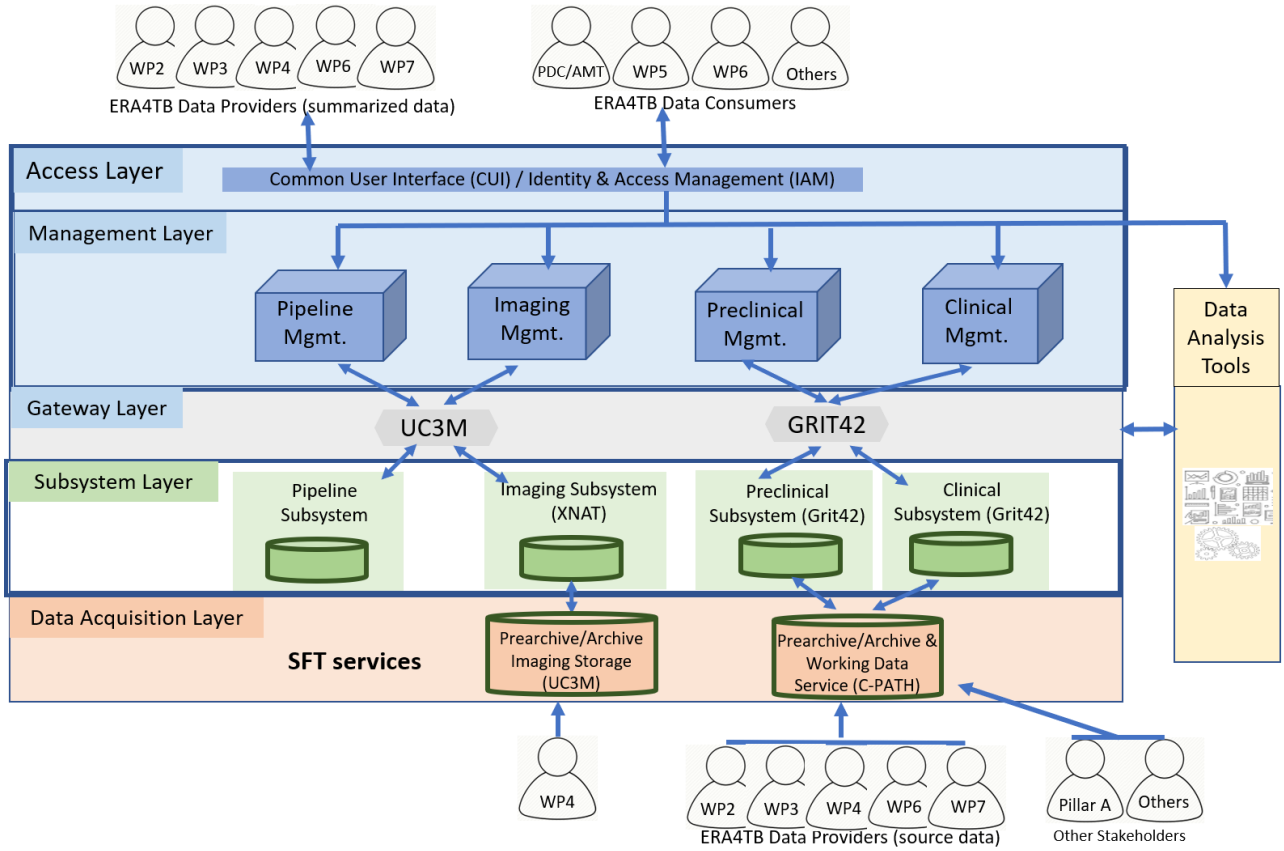

Figure 1. DDIM Architecture

DDIM is an evolving system that allows the integration or updating of new modules and technologies to respond to changing requirements of the ERA4TB data consumers and providers. The DDIM architecture considers the following layers:

1) Access
2) Management
3) Gateway
4) Subsystems
5) Data acquisition

A brief explanation of each layer will be given in the following sections.

2.1 Access layer

The function of this layer is to provide a common user interface (CUI) to the DDIM users, facilitating the authentication and access to the different involved subsystems. In a general view, DDIM users can be:

1) Data Consumers
2) Data Providers
3) Data Consumers/Providers

Data Consumers are those interested in obtaining and analysing pipeline, assays/experiments, and imagery information. Data Providers supply this information. Data providers and consumers are required to be authenticated by the DDIM portal. Data providers that supply data through external data subsystems will be validated by the corresponding subsystems.

In the DDIM portal users can visualize, without previous authentication, general information describing the ERA4TB project. However, users interested in visualizing specialized information or providing new information must be authenticated by the DDIM access layer.

2.1.1 User registration

The user registration process is depicted in Figure 2.

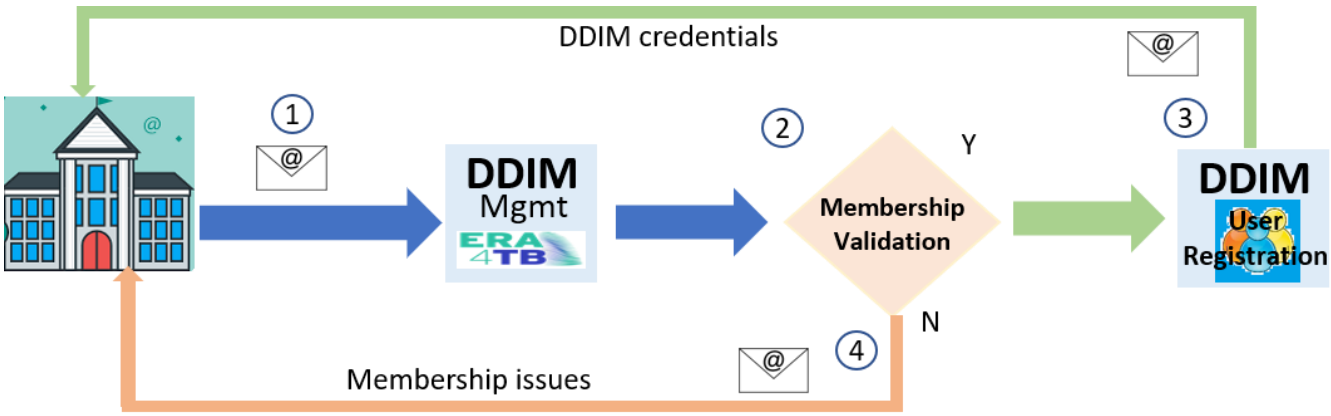

Figure 2. User registration process in DDIM

The tentative steps involved in the DDIM registration process are:

1) A representative of an institution, member of the ERA4TB consortium, sends an email to the DDIM manager with a request for registration in DDIM for one or more users that belong to that institution.
2) The DDIM manager will verify the correct ERA4TB membership of the interested institution. This validation includes a verification process of the type of roles (access permissions) that are requested for the users.
3) If membership validation and requested roles are correct, the involved users and institution representative will receive an email with their credentials (username and password) to access the DDIM system and subsystems.
4) If validation issues arise, they will be notified to the institution representative to resolve them and submit the request again.

Should a user needs changing its role or any profile data, a similar process as the mentioned above (steps 1 to 4) will be carried out, in which the involved institution will send to the DDIM manager a request for user profile update. The following table presents the user information that must be provided to the DDIM manager to carry out the registration process.

Attribute	Description
Name	Username (e.g., John)
Surname	User surname (e.g., Smith)
Institution name	Full institution name (e.g., University Carlos III of Madrid)
Institution acronym	Institution acronym (e.g., UC3M)
Email	User email (e.g., jsmith@uc3m.es)
Molecule/Compound	Molecule or compound of interest
Phase	Phase or phases of interest (e.g., In Vitro, In Vivo, FTIH)
Role	Tentative: See user roles section below. E.g., Experiment owner, Asset owner, Data modeler, etc.
Subsystem	Indicate the subsystem(s) that would be of interest: E.g., Preclinical, Clinical, Imaging, Pipeline, and Data analysis subsystems.

Table 1. User information required for registration

Table 2 shows a provisional list of user roles for the DDIM. Definitive roles are to be further defined by WP1 in conjunction with end user requirements from the various partners. Detailed permissions and scope for every role have to be given (e.g., Create, Read, Update and Delete).

Role	Description
Experiment Owner	Run pre-clinical and clinical experiments/ studies and may be looking for hypothesis-generating data or wish to identify gaps between data that exist and studies they may design.
Asset Owner	Owners of the molecules that have entered the drug progression pipeline in ERA4TB
Data Modeler	WP5 members looking at and/or analysing datasets across preclinical and clinical experiments to identify potential drug development tools (e.g. translational models and approaches, dose/dose regimen selection tools) or disease progression models that will support decision making. This user may also wish to access existing models or tools and apply these on their own data.
Consortia Member	Any member of the consortium who may wish to view data relating to the ERA4TB project
Data Manager	Members of WP1 responsible for data curation, quality control, transformation and the loading of data into the grit42 platforms

Table 2. User roles in DDIM

2.1.2 User authentication

Once a user is registered, the user in question can access private information in the DDIM according to the assigned role. A general description of the steps followed by a user during the identity and access management process is as shown in figure 3:

1) A user provides DDIM portal his credentials: user and password.

2) The DDIM-IAM module verifies user credentials in its user database.

    a) DDIM allows the authorized user to browse the portal information and navigate through the available DDIM management subsystems (e.g. pipeline mgmt., imaging mgmt., etc.). The management subsystems allow to visualize abstracted information that is locally available in DDIM or to redirect the user to the corresponding external subsystems when more detailed information is required.

    b) The access is denied for unauthorized user.

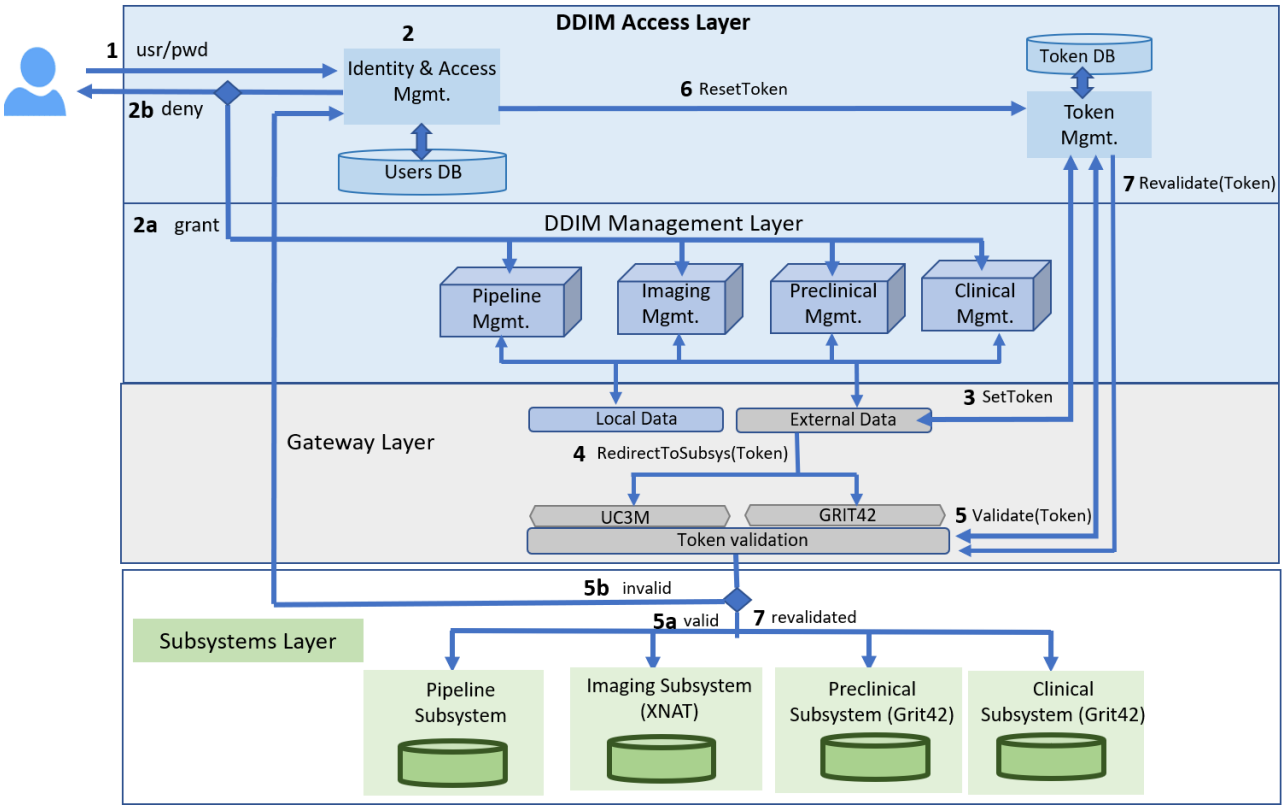

Figure 3. Identity and access management

3) Since DDIM acts a centralized authentication system for most of the subsystems, before a redirection to a subsystem (preclinical, clinical or imaging) occurs, a user's token is generated. This token allows subsystems to identify authorized user. This token is generated according to an open standard specification, Jason Web Token (JWT, IETF 7519). It is important to note that data acquisition subsystems (Secure File Transfer -SFT- services in Figure 1) are out of the authentication scope of the DDIM-IAM module.

4) User is redirected to a specific subsystem through the corresponding gateway connector. This redirection will include the user's token previously generated. The gateway connector is implemented by the subsystem provider, e.g., grit42 or UC3M.

5) It is expected that the token included in the subsystem redirection is valid. However, tokens have an expiration time, which means that for security reasons, subsystems must validate the token with DDIM from time to time. The corresponding subsystem's connector will call a validation REST function in DDIM, which validates if the user is still authorized and/or the token has not expired. An example on how to call this function is this:
https://ddim.era4tb.arcos.inf.uc3m.es/default/tk_validation?tk=XXX.YYY.ZZZ
Where XXX.YYY.ZZZ is the user's token whose structure is explained in next section.

   a) Valid users will be able to continue and accessing the corresponding subsystem.

   b) Unauthorized users will be redirected to the DDIM-IAM module, where user credentials must be provided.

6) User redirected from previous step will be validated by DDIM-AIM.

    a) For authorized users a new token is generated, and token revalidation process is executed (step 7).
    b) Access is denied to unauthorized users (step 2b).

7) Token revalidation allows user accessing to the corresponding subsystem. This is a similar function as defined in step 5.

Token structure

The token structure is based on the open standard specification, Jason Web Token (JWT, IETF 7519). JWT defines a compact and self-contained way for securely transmitting information between parties as a JSON object. This information can be verified and trusted because it is digitally signed using a secret (with the HMAC algorithm) or a public/private key pair using RSA or ECDSA. A common scenario for using JWS is for user authorization and information exchange.

JSON Web Tokens consist of three parts separated by dots (.), which are: Header, Payload and Signature. Therefore, a JWT typically looks like the following: xxxxx.yyyyy.zzzzz.

The header typically consists of two parts: the type of the token, which is JWT, and the signing algorithm being used, such as HMAC SHA256 or RSA. Example:

```
{
  "alg": "HS256",
  "typ": "JWT"
}
```

This part of the JWT structure is Base64Url encoded.

The payload contains the claims, which are statements about an entity (typically, the user) and additional data. For example, our first version of the token contains:

```
{
  "user": “username”,
  "rol": "privilegies_group",
  "exptime": “expiration date and time”
}
```

The username field refers to the name used to identify a valid user in DDIM. The rol field represents a set of privileges that the user has according to a group of users (roles) to which is assigned. exptime indicates the date and time the token expires. This field is currently defined as string in the following format: YYYY:MM:DD:HH:MM:SS

Where YYYY= year, MM=month, DD=day, HH= hour, MM=minutes, SS=seconds. This part of the JWT structure is also Base64Url encoded.

The last part of the token is the signature. To create the signature, you have to take the base64URL encoded header, the encoded payload, a secret, the algorithm specified in the header, and sign that.

For example, if you want to use the HMAC SHA256 algorithm, the signature will be created in the following way:

Signature =HMACSHA256(base64UrlEncode(header) + "." + base64UrlEncode(payload), secret)

The final structure will be: base64encodedHeader.base64encodedPayload.Signature

Nowadays, there exists free and open available libraries for encoding and decoding JWT tokens in different languages such as: Java, PHP, Python, etc.

Token validation function

DDIM provides the following REST function that can be called by supporting subsystems:

https://ddim.era4tb.arcos.inf.uc3m.es/default/tk_val?tk=XXX.YYY.ZZZ

When a subsystem calls this DDIM function, the following attribute-value pair is sent in the body of an HTTP response (in Jason format):

{“validation”:1}
or
{“validation”:0}

If response is {“validation”:1} means the token is valid, so the user can access the subsystem. If the response is {“validation”:0}, invalid token, the subsystem must invoke the following DDIM function:

https://ddim.era4tb.arcos.inf.uc3m.es/default/set_token

This DDIM function will take the user to the DDIM-IAM module to validate credentials again. For valid user, DDIM generates a new token that will be sent to the subsystem as the HTTP body response, as follows (Jason format, attribute:value pair):

{“token”:”XXX.YYY.ZZZ”}

if credentials are not valid the access will be denied.

2.2 Management layer

The DDIM system facilitates users to access the following modules:

1) Pipeline Management
2) Imaging Management
3) Preclinical Management
4) Clinical Management

These modules allow users to visualize, analyse and provide data associated to a compound-based pipeline, tuberculosis imagery, preclinical and clinical stages during the drugs development.

Data shown in these modules can be obtained from or provided to a local database, managed by DDIM, or extracted from external data subsystems associated to DDIM.

2.2.1 Redirection to subsystems

Users interested in managing detailed preclinical/clinical data and/or images are redirected to the corresponding subsystems, as explained in the identity and access management process (Step 4 of user authentication). For this end, DDIM calls a REST function implemented in the corresponding subsystem's gateway connector. This function call includes the user's JWT token that represents the current user. An example of this call to the Grit42's clinical subsystem would be:

https://clinical.era4tb.arcos.inf.uc3m.es/access?tk= XXX.YYY.ZZZ

Where the token XXX.YYY.ZZZ is the JWT token generated in Step 3 of user authentication. The subsystem will be able to decode the token, using a JWT library compatible with the IETF 7519. The typical JWT decoding function returns the payload of the token, which in this case contains the attributes user, rol and exptime, such as is shown in the example:

```
{
  "user": “user”
  "rol": "privilegies_group",
  "exptime": “expiration date and time”
}
```

These attribute-value pairs can be saved by the subsystem if required. If the subsystem only cares about the validation of the token, it will not be necessary for the subsystem to extract the payload from the JWT token. For security and control reasons, after receiving a call like this, the subsystem should execute the token validation function explained in the previous section. It is responsibility of the subsystem, check token validation from time to time.

2.3 Gateway layer

This layer abstracts the access to different data subsystems associated to the DDIM portal, e.g. Grit42. The function of this layer is to facilitate subsystem integration and reduce complexity. Every subsystem must offer DDIM a REST function that allow DDIM to redirect users to the corresponding subsystems, where a JWT token is included with user information. A general call to a subsystem would be:

https://subsystem.com/access?tk= XXX.YYY.ZZZ

Where tk is the JWT token provided by DDIM, which identifies the user interested in accessing data from the subsystem.

2.4 Subsystems layer

This layer encapsulates all data subsystems that will provide DDIM users with detailed information. It is a set of systems, applications and databases that will interact with DDIM by means of REST invocations.

The current available subsystems are:

1) grit42 preclinical subsystem
2) grit42 clinical subsystem
3) XNAT imaging subsystem

2.5 Data Acquisition layer

This is the lower layer of the DDIM architecture, which represents the data upload points for the preclinical, clinical, and imaging subsystems. Secure FTP servers are the main building blocks systems used in this layer. These servers play an intermediary role to facilitate communication between the data/image providers and the preclinical and imaging subsystems (e.g., grit42, XNAT).

Figure 4 and 5 illustrate how data and images flow from preclinical data providers (Figure 4) and image providers (Figure 5) to the corresponding subsystems.

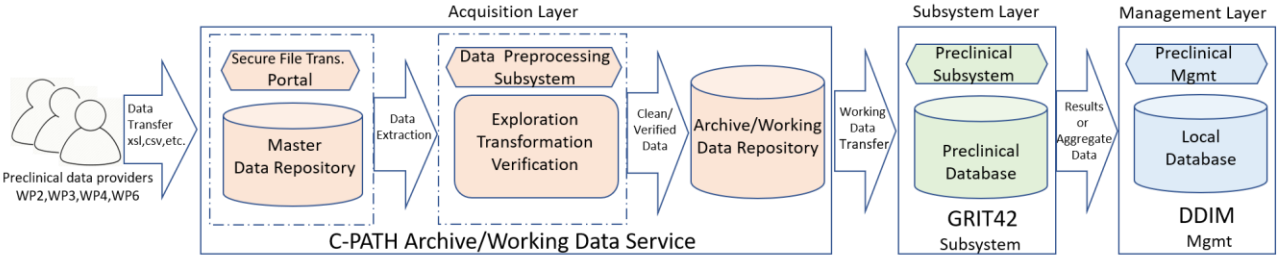

Figure 4. Preclinical data upload

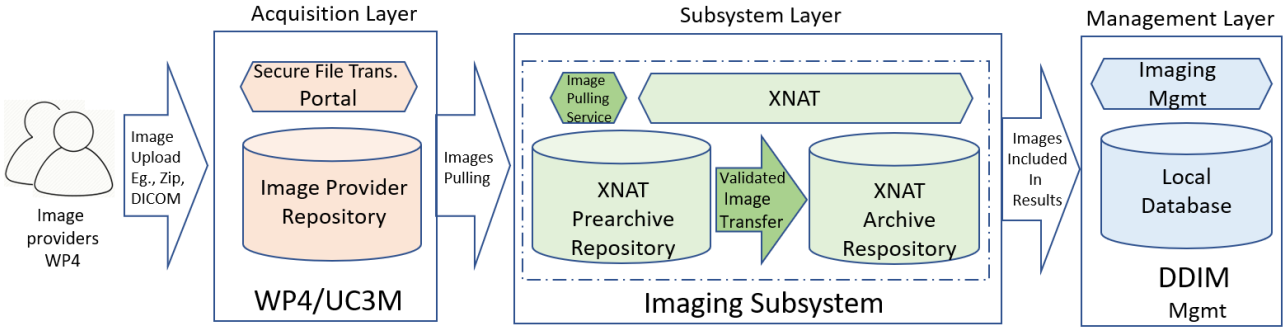

Figure 5. Images upload

Figure 4 shows how preclinical data is introduced into DDIM using a secure file transfer (SFT) portal. At this point, the data acquired from the data provider is stored in a Master and Working Data repositories managed by C-PATH. The C-PATH Data Collaboration Centre (DCC) team will apply various processes to the data, such as data exploration, quality control (QC) and standardisation/transformation. Once the relevant processes have been successfully applied, the data will be available in the Archive/Working Data Repository and loaded to the relevant grit42 instance which integrates this information as part of assays and/or experiments associated to a molecule or compound in the drug development process.

In Figure 5, it is expected that image providers have a secure FTP server, which represent the Image Provider Repository. In this repository, image providers can put the images that would like to send to the DDIM imaging subsystem (XNAT). Once the images arrive into the Image Provider Repository, an Image Pulling Service will get those images and store them in the Prearchive Repository of the Imaging Subsystem (XNAT). The images will be verified to determine if they do not have potential issues that could be against the image distribution restrictions in DDIM. If images are free of issues, they will be transferred into the XNAT's Archive Repository to be available to the DDIM imaging management subsystem if necessary. Images with distribution issues will be removed of the repository and this situation will be informed to the image provider.

It is expected that the images are compatible with the DICOM format and are compressed in a zip file. It is also expected that the images are included in a structure of directory that facilitates the identification of the compound/molecule, assay, and experiment that the images belong to. For example, if the file image0001.zip is received, it is expected that after uncompressing the zip file, the resulting directory structure and file will be similar to:

/compoun01/subject01/experimet01/image0001.dcm.

Where compound1, subject01, and experiment01 are the name of the molecule or compound, the experimentation object, and the experiment respectively to which this image should be associated.

Clinical data (WP7) will flow in a similar way to the preclinical data as shown in Figure 4. In addition to the data processes described for preclinical data, additional safeguards are applied to ensure that Personally Identifiable Information (PII) is not delivered to the DDIM.

3. Deployment

This section briefly describes the UC3M proposal for the deployment of DDIM. The different layers and modules of the architecture can be encapsulated in virtual containers using the Docker platform. Figure 6 shows how the DDIM portal can access data from the involved subsystems.

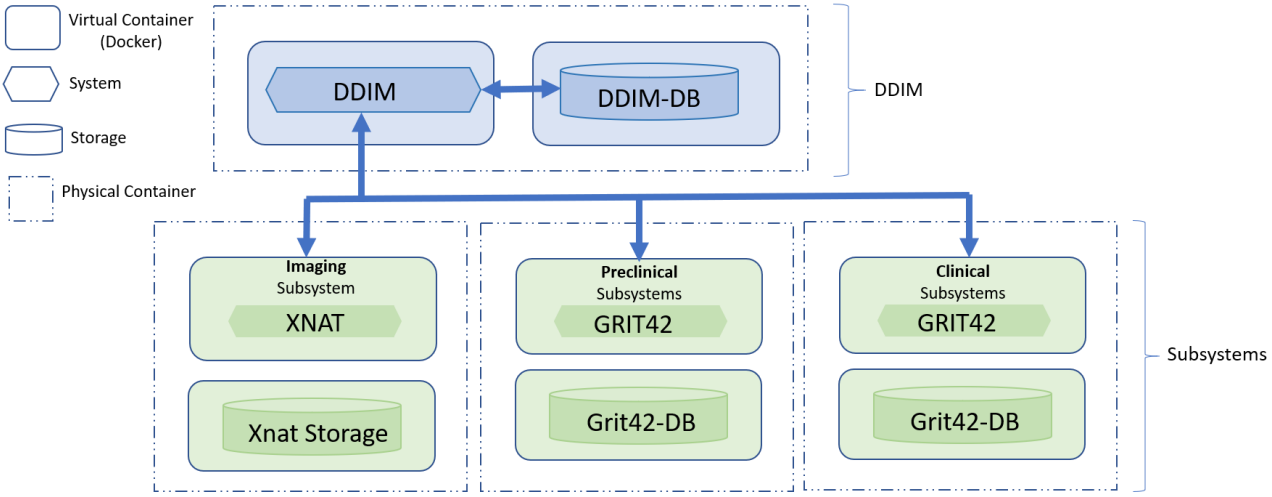

Figure 6. DDIM deployment

To facilitate deployment and migration, DDIM and data subsystems will be executed in a virtualized platform (Docker). Subsystems and databases can be hosted in different containers. Communication between DDIM and subsystems is based on REST invocations. Virtual containers[1] are hosted by physical containers (real servers).

4. Data Flow

This section provides a generic preliminary description on how compound/molecule data is received, processed, and presented by the subsystems involved in the DDIM system.

[1] https://www.docker.com/resources/what-container

We have divided the data flow into three main stages: Data Input, Data Processing, and Data Presentation. These 3 stages are supported by a Data Storage service as shown in Figure 7.

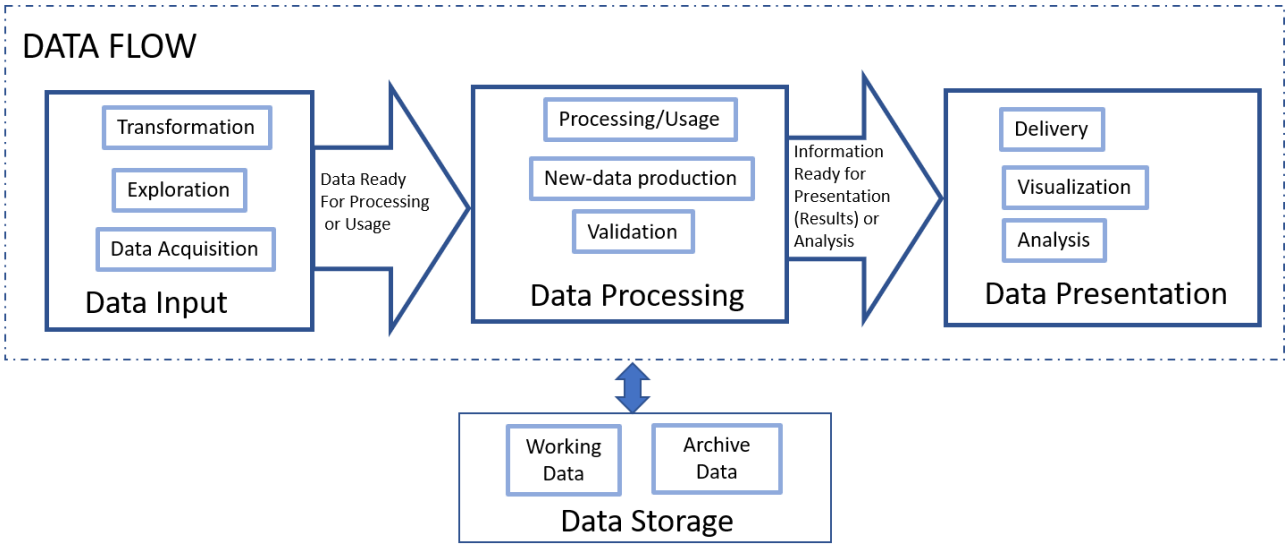

Figure 7. General data flow

Data Input. During this stage data suppliers provide data associated to a Compound/Molecule that will be used as part of the drug development process. The template for this data is defined according to the specification given by the preclinical and clinical external subsystems, which support the DDIM system, and it is out of the scope of this document. In this stage, the data are treated with different processes such as data quality control (QC) and/or data transformation to avoid error propagation in data, when manipulating by the preclinical and clinical subsystems. The output of the Data Input stage are data ready to be used in experiments and assays required in the involved subsystems (preclinical and clinical).

Data Processing. In this stage, the data produced by the Input Data stage is used as part of the information included in experiments and assays carried out with the Compounds/Molecules during the drug development process. This information is reported using the preclinical and clinical subsystems. New data can be produced from these experiments/assays. The summary of the results obtained from the experiments associated to one or more assays represents the output produced by the Data Processing stage. This stage also considers an information validation process, where the results obtained are verified to discard the presence of restriction issues, e.g., Protected Health Information (PHI), Personally Identifiable Information (PII), ambiguous values, etc.

Data Presentation. This is the last stage of the Data Flow. Information produced by the Data Processing stage can be used for different purposes in the Data Presentation stage. Examples of usages are visualization and analysis, by using another DDIM subsystem, or only for delivery of final reports.

Data Storage. This is a federated storage service that complement the data flow stages as shown in Figure 5. DDIM subsystems involved in all of the data flow stages will be able to share information from their corresponding database systems.

4.1. Data Templates for Reporting Results

The results obtained from the assays carried out on a Compound/Molecule during a phase of the pipeline should be reported following specific data templates. To achieve this end, in this section we provide the following information:

a) Information Entities. Represent the description of the information entities, and their relationships, which should be considered in the results obtained from the Data Processing stage.

b) Data Templates. Describe the structure of the reports that contain the concrete results obtained from the implemented assays.

4.1.1 Information Entities

Information entities represent the basic elements of information that should be considered when reporting results. Figure 8 shows an entity-relationship diagram that includes the main entities involved in the expected results generated by the preclinical and clinical assays, which in turn are managed by the corresponding DDIM subsystems. The entity-relationship diagram describes the relationships among the most relevant information entities considered in the pipeline. The following are the descriptions of these relationships:

a) Participant-Compound relationship. A Compound/Molecule (C/M) belongs to at least one owner (Participant entity). This part of the relationship is denoted by (1,n). The other part of the relationship states that a participant (owner of a C/M) can have 1 or more C/Ms in the pipeline, which is denoted by (1,m).

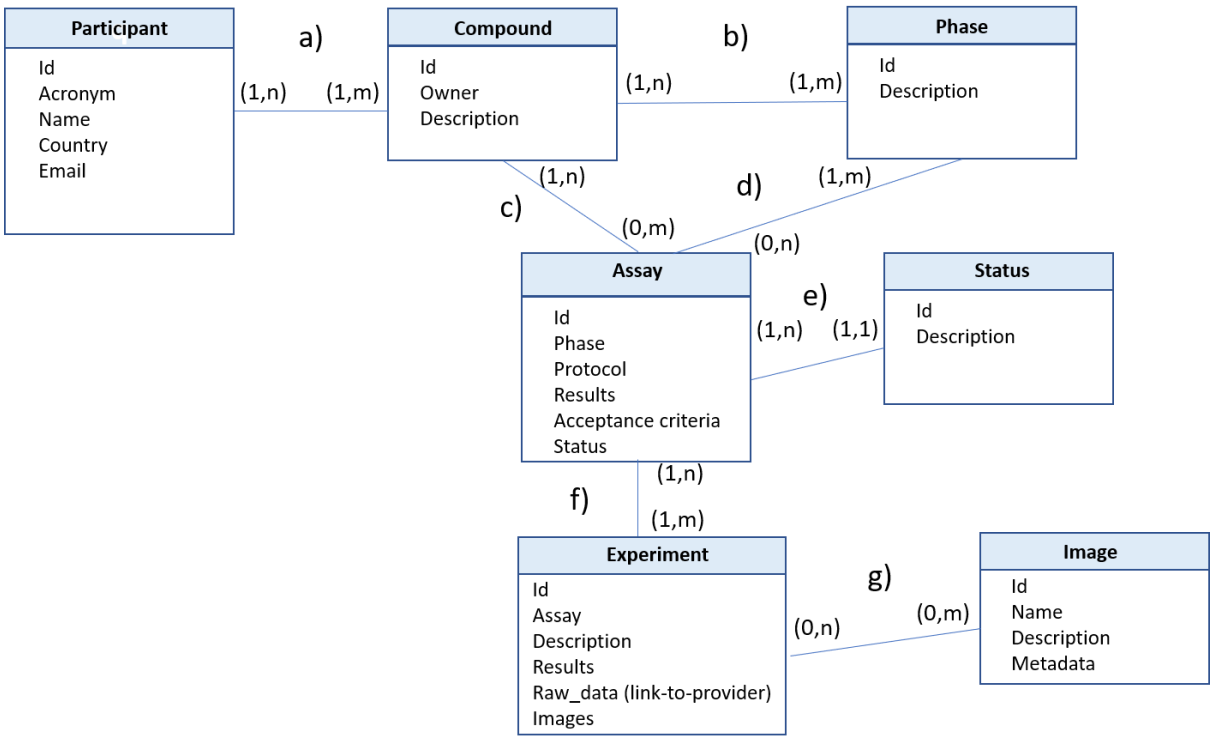

Figure 8. Information entities that should be considered in the results generated from preclinical/clinical essays.

a) Compound-Phase relationship. A C/M can be used in one or more phases of the drug development process (1,m). In a phase of the pipeline is possible to manage 1 or more C/Ms (1,n). The phases considered in this version of the document are described in the next section (see Phase entity description).

b) Compound-Assay relationship. A C/M could be or not currently tested in one or more assays (0,m). Every assay needs to include at least one C/M (1,n).

c) Assay-Phase relationship. An assay must be associated to one or more phases of the pipeline (1,m). In a phase of the pipeline could be zero or more active assays (0,n).

d) Assay-Status relationship. Results of an assay have a status assigned (1,1). Different assays can have the same status (1,n). The list of status that can be assigned to assay is given in the next Section (see Status entity description).

e) Assay-Experiment relationship. An assay can has 1 or more experiments associated to itself(1,m). One experiment can be applied in different assays (1,n).

f) Experiment-Image relationship. An experiment can include zero or more images (0,m). One image can be included in zero or more experiments (0,n).

Next, a description of each information entity is provided.

Participant. It represents an institution, organization or individual that provides a compound or molecule (i.e., the owner), intended to be part of the drug development process that will be carried out by the ERA4TB Consortium members.

Table 3 shows the information elements that describe every attribute of the Participant entity. The Name column indicates the internal name this attribute has for the DDIM system. The Label column shows how end users will identify this attribute. Description/Comments. It provides general information about the attribute. Format indicates the internal data type used for this attribute. The R-C column shows if the value of the attribute is required (R) or conditional (C).

Name	Label	Description/ Comments	Format	R-C
Id	Compound Owner	The owner (institution, organization or individual) of the Compound/Molecule (C/M) to be used during the drug development process.	INT	R
Acronym	Acronym	Acronym of the organization if any.	VARCHAR	C
Name	Name	Official name of the C/M owner	VARCHAR	R
Country	Country	Country where the C/M owner officially resides (ISO 3166-1 alpha-3 standard).	VARCHAR	R
Email	Email	Official email of the C/M owner (contact person)	VARCHAR	R

Table 3. Attributes of Participant entity

Compound. It represents a Compound or Molecule intended to be part of the drug development process carried out by the ERA4TB Consortium members. Table 4 shows the information elements that describe the attributes of the Compound entity.

Name	Label	Description/ Comments	Format	R-C
Id	Compound-Molecule Id	Compound/Molecule (C/M) to be used during the drug development process.	VARCHAR	R
Owner	Owner	A valid reference to an Id in the Participant entity	INT	R
Description	Description	A concrete description of the C/M	VARCHAR	R

Table 4. Attributes of Compound entity

Phase. It represents the phases that are considered in the current version of the pipeline. It is a catalogue entity that includes the following values:

a) Precandidate Entry Criteria AU
b) Preclinical Candidate Development
c) First Time in Human (FTIH)

Table 5 shows the information elements that describe the attributes of the Phase entity.

Name	Label	Description/ Comments	Format	R-C
Id	Pipeline Phase	One of the phases considered in the pipeline. For this version of the document are: Precandidate Entry Criteria AU; Preclinical Candidate Development; and First Time in Human (FTIH)	VARCHAR	R
Description	Description	A concrete description of the C/M	VARCHAR	R

Table 5. Attributes of Phase entity

Assay. It describes the protocol and results of a specific assay. Table 6 shows the information elements that describe the attributes of the Assay entity.

Name	Label	Description/ Comments	Format	R-C
Id	Assay Identifier	Unique identifier for an assay	VARCHAR	R
Phase	Phase	Reference to a valid Phase Id to which this assay is associated to.	VARCHAR	R
Protocol	Protocol	Text or PDF file that describes the assay.	File	R
Results	Results	Excel file with the results of the assay.	File	R
Threshold	Acceptance Criteria	Values that indicate if a C/M goes on or not.	INTARRAY	R
Status	Status	Reference to a valid status that can be assigned to an assay.	INT	R

Table 6. Attributes of Assay entity

Status. It describes the status of an assay. The following are the considered status for this version of the document:

0. Study completed and there is a risk that needs to be mitigated
1. Study not done
2. Study ongoing
3. Study completed and does not preclude further development of molecule

Table 7 shows the information elements that describe the attributes of the Status entity.

Name	Label	Description/ Comments	Format	R-C
Id	Status	Unique identifier for a status	INT	R
Description	Description	Description of one of the possible status for an assay: (0) Study completed and there is a risk that needs to be mitigated, (1) Study not done, (2) Study ongoing, (3) Study completed and does not preclude further development of molecule	VARCHAR	R

Table 7. Attributes of Status entity

Experiment. It provides a description of an experiment associated to an assay. Table 8 shows the information elements that describe the attributes of the Experiment entity.

Name	Label	Description/ Comments	Format	R-C
Id	Experiment Identifier	Unique identifier for an experiment	VARCHAR	R
Assay	Assay Id	Reference to a valid Assay Id to which this experiment is associated to.	VARCHAR	R
Description	Description	Description of the experiment	VARCHAR	R
Results	Results	Excel file with the results of the experiment.	File	R
Data	Reference to data	It contains a URL to get access to data provided by a DDIM external subsystem, which are associated to this experiment.	URL	R
Images	Reference to complementary images	It contains a URL to images provided by a DDIM external subsystem associated to this experiment.	URL	C

Table 8. Attributes of Experiment entity

Image. It provides information about the images associated to experiments or assays that are carried out on a C/M. Table 9 shows the information elements that describe the attributes of the Image entity.

Name	Label	Description/ Comments	Format	R-C
Id	Image ID	Unique identifier for an image	VARCHAR	R
Name	Name	Mnemonic name of an image associated to an assay or experiment	VARCHAR	R
Description	Description	Description of the image	VARCHAR	R
Metadata	Metadata	Meta information that would be included for image description and indexing (to be defined).	VARCHAR	R

Table 9. Attributes of Image entity

4.1.2 Data Templates

This section provides the template structures for gathering summarized information of the results obtained from the Data Processing stage. As first approach, this template is generated following the guidelines given by ERA4TB project leader (GSK) and coordinator (UC3M). Three development stages are considered in the pipeline for every compound/molecule:

a) Precandidate Entry Criteria AU

b) Preclinical Candidate Development

c) First Time in Human (FTIH)

Fields within each template are designed to capture the concrete and relevant result of the corresponding assay. Detailed and additional information is accessed through hyperlinks to external DDIM subsystems that provide the complementary information of the experiments and the data.

As a first approach, we have defined a homogeneous template for every stage. Data required to be reported in the three templates are shown in Table 10.

Compound/molecule ID
Assay ID
Criterion
Assay status
Results
Acceptance Criteria
Supporting information

Table 10. Data template for reporting results

The following is the description of the attributes of this template.

Compound/molecule ID:

Description: Identification number/code for a compound or molecule involved in the assay.
Field type: String.
Tentative format: CCCC-DDD, where C: Char, D: Digit.

Assay ID:

Description: Identification number/code for the study/assay.
Field type: String.
Tentative format: CCCC-DDD-PP-TS-S , where CCCC-DDDD compound Id; PP: Digit representing phase/stage (00: Precandidate Entry, 10: Preclinical
Candidate Development, 20: FTIH phase I, 21: FTIH phase 2,……, 29: FTIH phase 9; TS: Type of study (number from 00 to 99); S: Digit representing study status.

Criterion:

Description: Concise description of the assay.
Field type: Text.

Assay status:

Description: Status of the current assay. Where possible values are 0, 1, 2 and 3 (they can also have color representation):

0. Study not done (orange).
1. Study ongoing (yellow).
2. Study completed and there is a risk that needs to be mitigated (red).
3. Study completed and do not preclude further development (green).

Field type: Int.

Results:

Description: Specific and concrete values resulting of the assay.
Field type: Text.

Acceptance Criteria:

Description: A list of minimum/maximum expected values (the measure units will be given) that determine if the procedure continues (go) or not (no-go).
Field type: Array of float.

Supporting information (data):

Description: Link to external data sources or experiments that support the obtained results. The structure of this information to be defined.
Field type: URL.

Example of reports of results for different pipeline development stages are given in Annex 1, 2 and 3. Most information provided in these reports is fictitious.

Annex 1. Example of results for Precandidate Entry Criteria AU phase.

compound_id	assay_id	entry_criterion	Status	Results	Acceptance Criteria	Supporting Information
MOL-001	MOL-001-00-00-3	Program MOA (relative to current TB drugs)	**3**	ATP synthase inhibitor with potential for larger safety margin than BDQ	ex: Novel Not novel but compound is significant improvement over currently available (precise which one and to what extent)	http://grit42.com/data
MOL-001	MOL-001-00-01-3	MIC - aerobic (H37Rv M. tb)	**3**	< 0.004 - 0.01 ugml		http://grit42.com/data
MOL-001	MOL-001-00-02-3	MIC in 10% serum - aerobic	**3**	with 50% human serum - MIC = 0.29 ugml. However efficacy seen in mice despite simiar shift with mouse serum.	ex: <3 fold shift in MIC < 10 fold shift in MIC < xx fold shift in MIC	http://grit42.com/data
MOL-001	MOL-001-00-03-3	MIC - anaerobic (90% CFU reduction)	**3**	0.007 - 0.07 ugml (LORA)	ex: <10 fold shift in MIC < 30 fold shift < xx fold shift none if not mechanistically expected	http://grit42.com/data
MOL-001	MOL-001-00-04-3	MBC - aerobic (H37Rv M. tb)	**3**	0.03 – 0.1 ugml after 14d	ex: Bactericidal Bacteriostatic	http://grit42.com/data
MOL-001	MOL-001-00-05-3	MIC in macrophage (in 10% serum)	**3**	2 log CFU reduction in 7d at 0.1 ugml in J774 cells against Erdman	ex: Same as MIC in 10% serum None …	http://grit42.com/data
MOL-001	MOL-001-00-06-3	Target genetics	**3**	Essential in vitro and in vivo (in animals)	Essential in vitro and in vivo (in animals) Essential only in vivo (in animals)	http://grit42.com/data
MOL-001	MOL-001-00-07-3	Activity on target (if applicable)	**3**	Cross-resistance with BDQ on ATP synthase mutants.	ex: < 10 nM < 100 nM < xx nM	http://grit42.com/data
MOL-001	MOL-001-00-26-3	Drug transporter Pgp	**1**	Weak Pgp substrate	ex: not a Pgp substrate or inhibitor BA/AB ratio <5 if substrate…	
MOL-001	MOL-001-00-27-1	hERG inhibition	1	28.6uM (nonGLP)	ex: > 30 uM > 10 uM > xx uM	
MOL-001	MOL-001-00-28-1	Receptor/Enzyme/Channel Off-taget binding	1	L-type Ca channel Na channel MC1 receptor norepinephrine transporter	ex: No hit (<25% inhibition at 10 uM)	

Annex 2. Example of results for Preclinical Candidate Development phase.

compound_id	assay_id	entry_criterion	status	Results	Acceptance criteria	Supporting Information
MOL-001	MOL-001-10-00-3	Efficacy in Combination study Add Combos corresponding experiment and results	**3**	With pretomanid and linezolid - similar efficacy to BDQ at half dose; superior efficacy (clearance rate and treatment-shortening) at same dose (25mpk)	ex: at xx mg/kg (Therapeutix Index> xx fold) CFU clearance in a mouse relapse model with 3 month regimen in combination with compound partner B (xx mg/kg) and C (xx mg/kg). Intratracheal infection with 10xx CFUs and treatment by oral gavage (once a day BD ...)	http://grit42.com/data
MOL-001	MOL-001-10-01-3	NOAEL determined in Rats and higher species 14 day dose ranging toxicity	**3**	Rat NOAEL 40mpk; Dog NOAEL 60mpk (exposures >4-15 fold over predicted efficacious)	ex: >10 fold over efficacious exposure >3 fold over efficacious exposure …	http://grit42.com/data
MOL-001	MOL-001-10-02-3	Cross-Functional Development Plan Established	**3**	ex: An integrated product development plan to achieve impact in the populations of interest has been articulated. Risks are known and agreed to be manageable or an integrated product development plan to achieve impact in the populations of interest has been	ex: An integrated product development plan to achieve impact in the populations of interest has been articulated. Risks are known and agreed to be manageable or an integrated product development plan to achieve impact in the populations of interest has been	http://grit42.com/data
MOL-001	MOL-001-10-04-3	Initial drug substance characterization completed	**3**			http://grit42.com/data
MOL-001	MOL-001-10-05-3	solubility (pKa Log P/Log D)	**3**	pKa 8.33; LogP 4.64; LogD 4.05 (pH7.4)		http://grit42.com/data
MOL-001	MOL-001-10-06-3	permeability (Caco2 permeability coefficient)	**3**	Low Papp (see previous tab)		http://grit42.com/data
MOL-001	MOL-001-10-07-3	BCS classification	**3**	II		
MOL-001	MOL-001-10-20-1	PK in rodent and non rodent > xx times PD driver precise model (monotherapy acute or chronic or in combinations)	**1**	No - PKPD complex in mice due to active metabolites. Not yet complete. Plan to model and complete study when Phase I data available to inform parent/metabolite ratio	predicted human dose < 1g/d	

Annex 3. Example of results for FTIH phase.

compound_id	study_id	entry_criterion	status	Results	Acceptance criteria	Supporting Information
MOL-001	MOL-001-20-00-3	GLP Rodent 28d oral toxicity study	3	NOAEL 40mpk AUC 0-24>10fold predicted efficacious	detail possible findings	http://grit42.com/data
MOL-001	MOL-001-20-01-3	GLP Non-Rodent 28d oral toxicity study	3	Dog LOAEL 20mpk (AUC 0-24>13 fold predicted efficacious). NOAEL not identified. Findings at LOAEL min-mild reversible hepato-biliary phospholipidosis gastic mucosa.	detail possible findings	http://grit42.com/data
MOL-001	MOL-001-20-02-3	GLP Rat respiratory safety	3	No findings	ex: >10 fold >30 fold exposure	http://grit42.com/data
MOL-001	MOL-001-20-03-3	GLP Non-Rodent combined Resp/CV	3		detail possible findings	http://grit42.com/data
MOL-001	MOL-001-20-04-3	GLP Ames test (with and without metabolic activation) and MLA	3	Negative	ex: positive/negative	http://grit42.com/data
MOL-001	MOL-001-20-05-3	GLP micronucleus	3	Negative	ex: positive/negative	http://grit42.com/data
MOL-001	MOL-001-20-06-3	GLP in vivo micronucleus	3	Negative	ex: positive/negative	http://grit42.com/data
MOL-001	MOL-001-20-07-3	GLP CV study	3	No effects in 12day escalating dose conscious dog telemetry study Cmax >10X predicted efficacious	ex: >10 fold over efficacious Cmax	
MOL-001	MOL-001-20-20-1	GLP Rodent Irwin test	3	No findings	detail possible findings	