DO→IT

D2.1 Interim guidance document for existing BD4BO projects on approach proposed by the Distributed Data Network working group

IMI2 Project ID – DO->IT

Big Data for Better Outcomes, Policy Innovation and Healthcare System Transformation

WP2 – Knowledge Integration and Management

Lead contributor	Michel Van Speybroeck – (23 – Janssen)
	mvspeybr@its.jnj.com
Other contributors	Michael Arend – (15 – Bayer)
	Jeremy Hayter – (25 – Pfizer)
	Joachim Marti – (8 – Imperial College London)
	Anthony Rowe – (23-Janssen)
	Jerry Lanfear – (25 – Pfizer)

Due date	01 – May - 2017
Delivery date	03 – May - 2017
Deliverable type	R[1]
Dissemination level	PU

[1] **Use one of the following codes:**
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc.

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116055. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA

Description of Work	Version	Date
	V 1.2	3 May 2017

Document History

Version	Date	Description
V0.1	18 Apr 2017	First Draft
V1.0	25 Apr 2017	Reviewed version
V1.1	03 May 2017	Second revision
V1.2	04 May 2017	Final
V1.3	14 Sept 2017	Revised version

Contents

Introduction

The BD4BO initiative is a second-generation IMI program that aims to leverage “big data” to support the advancement of health outcomes research. To ensure that IMI resources committed to this program are spent appropriately and not wasted on duplicative work, it is critical that BD4BO projects re-use existing technological or governance solutions to the maximum extent possible, acknowledging this is fast paced field where technological advancements can and should be incorporated as appropriate. To that end, this subchapter summarises a set of operational and technology recommendations that encapsulate the lessons learnt from earlier IMI Data and Knowledge activities that have taken place within the IMI Framework.
New BD4BO projects should adhere to these recommendations as part of the development of their technology development and where deviating from them, justify their decisions in the development of the full project proposal.
There are three principles that underlie these recommendations:

- Data Management Practices in large research consortia are not a purely technical concern. They require new working practices between data providers and data consumers. Applying existing best working practices is critical to ensure that mistakes from earlier projects are not repeated.

- Data Management Technology has received considerable funding support from both EFPIA and Public funders in IMI, FP7 and H2020. To ensure best value for the European tax payer it is preferred that where technology is available and open for reuse, it is reused. Applying new technology for achieving the same purpose is discouraged.

- The BD4BO program has been developed as a strategic program that will enhance the overall competitiveness of health outcomes research. As such individual projects within the program that develop data management solutions that do not add to the overall value of the BD4BO will not contribute to this transformative strategic objective.

In combination, these three principles create a strong argument for new BD4BO projects to use existing technologies and governance solutions and not build new solutions, unless absolutely required by the outcomes questions that form part of the study.

Data Management in Research Consortia

From a data management perspective, it is helpful to categorise the stakeholders and the data access scenarios that are being used.
With respect to a given data set, the categories of functional participants are (note that these categories are not mutually exclusive):

- **Data Subject** - The individual to which the data relates. (note that the term data subject here refers to the functional definition and not necessarily to the definition as per privacy regulation i.e. for this further discussion data related to a particular data subject might be fully de-identified or not)

- **Data controller** – The entity that manages a dataset and controls access and content of the dataset. They may be a clinical site, a research laboratory or other data producing facility. Data owner and data controller might be one and the same organization while in other cases (e.g. regional health databases or primary care databases) it will be different organizations or individuals

- **Data co-ordinators** – These are member of the consortium who are typically data management professionals that seek to enable the consortium by supporting the harmonisation and integration of the study data on behalf of the consortium.

- **Data Consumers** – These are all members (including associate members) of the consortium who would want to use the data either individually or part of an integrated consortium wide data set to answer the research questions that are the objectives of the consortium.

With respect to data access scenario categories – these fall into:

- **Use of data collected during the project** - This scenario occurs where the consortium sponsors a new research study to collect real world data. In the context of the IMI project data generated during the time of the project is considered foreground IP. The patient consent and ethical review are consistent across the project and data management tools are all generated by the consortium. For example, a novel biomarker discovery cohort study (IMI Examples include U-BIOPRED & Oncotrack)

- **Use of data collected prior (or outside) the project** – The scenario occurs where the consortium aims to pool a number of background datasets provided by consortium members for either a secondary use to its initial purpose or an extended use to in initial purpose. In the context of and IMI project this data is considered background IP and owned by the data owner who is responsible to ensure its use is compliant with both the patient consent and ethical review boards who approve its use. Examples include EMIF and Predict-TB)

With respect to data types, three main types can be distinguished:

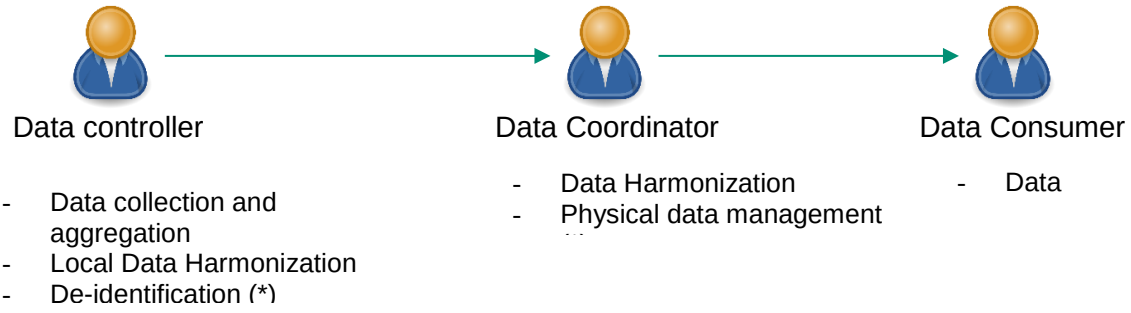
- Directly identifiable data: the study participant can easily be identified by means of name, address, contacts details, or other direct identifier.
- Pseudonymised data: identification of the study participant is not possible without additional information which is held separately from the study data (e.g., key coded data). Identification is possible using reasonable means.
- Anonymous data: the data cannot reasonably be associated to a particular identifiable study participant.

Data protection laws and the GDPR only apply to the first two categories of data and not to anonymous data. Where research objectives can be met with anonymous data, anonymous data should be used. If not, studies should use pseudonymised data.

Logical and Physical Data Flow

The logical data flow describes how and by which organizational data are transformed from their 'raw' or source format into analysis results. A physical flow indicates how data are physically transferred and where data are stored.

In relation to the physical data flow, the implementation is dependent upon the specific use case in the respective project but the following guidelines can be applied:



(*) As applicable, depending upon use case

Figure 2: Logical flow of data across different stakeholders

1) Disclosure of individual patient level data (IPD) should be kept to an absolute minimum, especially if it relates to directly identifiable data
2) Pseudonymisation/anonymisation should occur as close as possible to the original source – to the extent that can be accommodated by the research questions
3) Data controllers should keep control over the access to their data - especially for IPD – irrespective of the place where these data are stored

Working Practice Recommendations

In the context of the BD4BO collection, it could be anticipated that a significant portion of the data used will initially come by accessing data from existing real-world data resources, such as electronic health records, case management systems or patient registries.
A useful reference in this scenario is the "Code of Practice for Secondary Use of Medical Research Data" [21, 24].
A key consideration for a consortium in making secondary use of the data is how they plan to pool pseudonymised/anonymised individual patient data (IPD) for the purposes of enabling the most flexible data analysis approaches.
An approach that has been implemented with success is to keep access to IPD only at the data controller level and "bring the analysis to the data" instead of pooling data together in a data warehouse to subsequently perform the analysis on the pooled data. To achieve the necessary scalability, this approach consists of the following steps:

1. A consortium will agree on a common data model
2. Each data controller will transform their data into a local version of the common data model

3. Research questions will be developed and approved by the respective Internal Review Board (IRB) and an “analysis script” - a computer program that assumes the common data model - is developed.
4. The IPD element of the analysis script is sent to each data controller who then runs this on their common data model. Summarised patient data are returned from each controller to the data co-ordinator who can then pool these data sets for continued analysis

This federated model with only data controllers processing IPD and other consortium members processing summarised data has been demonstrated as the generally best approach to facilitate analysis across distributed network of data providers, by balancing the risk of data misuse against the need of technical innovation. It has been implemented by the European and US based projects: FP7 EU-ADR, IMI EMIF, OMOP, OHDSI and the FDA Sentinel project.

This model addresses a lot of the key concerns around sharing of IPD but might be incapable of addressing certain research questions (e.g. when requiring pooled data sets for statistical analysis). In these situation, a physical pooling of pseudonymised/anonymised IPD is the only viable approach. A central third party (the data coordinator) can assume here responsibility for data harmonization and provide pooled data sets to the community of analysts (statisticians, informaticians and machine learners).

In this case, the data harmonisation is the responsibility of the network of data controllers within the consortium. Each controller will need to ensure that the processing of their data is compliant with:

- Local privacy regulations and is suitably pseudonymised/anonymised for scientific purposes (i.e., context-specific small risk of re-identification)
- The ethical governance and patient consent where required. Each use of the data may need to be reviewed by an IRB or data access committee
- The security practices and management of the third-party data processing capability are sufficiently mature that they can be audited.

Preferred System Recommendations

The recommended working practices have implications for the type of Data Management Technology that is selected and deployed:

- It should support the recommended operating model
- It should support the common harmonisation standards that are required to deliver outcomes research
- It should support the strategic aims of the BD4BO program by being aligned to the other technology in this space

A useful introduction on considerations about data standards is the eTRIKS Standards Starter Pack [25]. This has been compiled by both clinical standards authority (CDISC) and the biomedical standards community (ISA Tab organisation)

With these conditions in mind the following overview should be used to help identify technology platforms complimentary to existing IMI investments and practice

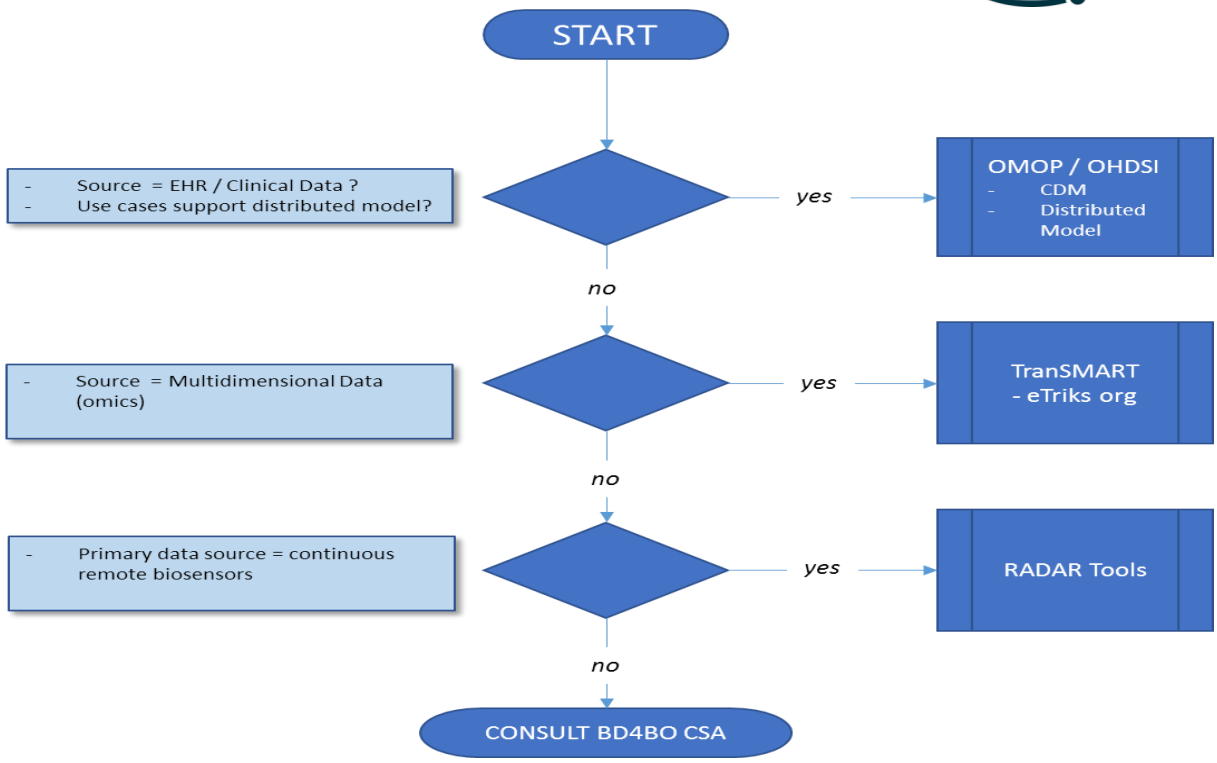


Figure 3: Technology selection overview to ensure compatibility with existing technology investments of the IMI. Depending on the architecture of the project is may be justifiable to select a combination of methods

OHDSI CDM & Tools

The OHDSI Project [26] is an approach and series of tools built for using healthcare data to support outcomes research. It is a successor to the FNIH OMOP project where the definition of the Common Data Model was defined for healthcare research that enabled the approach to processing IPD at a controller's site, but enable analytics to be run over all controller's in the network.

OHDSI tools enable a common technology standard that has been adapted by both EMIF and EU-ADR projects as the basis of enabling larger scale research networks than previously thought.

OHDSI tools are open source so do not require software licensing fees and EU SME's provide technology support to help projects leverage these tools.

eTRIKS/TranSMART

The TranSMART project [27] is an open source technology for providing knowledge management tools that has been most successful in prospective biomarker cohorts. It is a patient centric repository that captures patient data.

IMI eTRIKS project [28] has supported the use of tranSMART in over 30 research consortia. As a consortium they are willing to advice any IMI project on all aspects of data management technology that can be applied to translation research. From advice to training in how to deploy and use tranSMART in a project to sustainable support hosting translational research project data via the Luxembourg Elixir Node.

RADAR Tools

The IMI program RADAR is a sister to the BD4BO program that aims to better develop and validate the science of using continuous remote measuring technology such as wearable devices in Depression, Multiple Sclerosis and Epilepsy with other topics such as Alzheimer's planned in the future.

The RADAR consortium leverages technology such as purple robot [29], openSMILE [30] and tranSMART for knowledge management in this emerging area of science. They can be contacted via the IMI project office.

Considerations for projects

As mentioned in the introduction, these recommendations have been made on the basis of practical experience in health outcomes research in European consortia. They aim to balance the risk to data controllers with the ability to drive innovative research, to ensure that investments are complimentary and competitive, and to ensure a transformative future for health outcomes research in Europe across all disease area.

It is recommended that the discussion of data management is driven in partnership with data controllers of background sources. This will increase significantly the chance of user acceptance and generally accommodate a better use than defining the data management approach in isolation.

Data controllers will need allocation of sufficient resources in a project to take on the following responsibilities (who may be the projects data coordinators):

- Anonymise and transform their data for analysis, in preference by the common data model outlined
- If IPD is processed locally – resource to run local analytical tasks when received from the project
- If IPD can be pooled by a third-party data co-ordination site to audit and validate the technology environment.

This may require projects to reflect these recommendations in their scope / budgets.

Future Evolutions

It is recognized that the field of big data and outcomes research is in full development. The technico-legal aspects therefore need to be monitored closely when deploying solutions for new initiatives. Two relevant evolutions include:

- EU General Data Protection Regulation [31]: The upcoming change in data protection regulation will come into effect in May, 2018 and relates to the processing of personal data. Although – to the extent known- the above text is written with the current understanding of the upcoming GDPR regulation, specific elements
- Blockchain technology is a distributed ledger technology and is mostly known for its association with the digital currency Bitcoin but it has potential to be applied as well in healthcare [32]. It may pose some issues from a GDPR viewpoint since no data are deleted.

Summary

The above recommendations should be the basis for the default strategy for BD4BO initiatives. Where possible the proposed technologies should be applied - primarily to ensure IMI investments in BD4B0 are focused on defining health outcomes research and not on development of novel underlying technology. At the same time, technological evolutions, the emergence of novel data types or analysis methods or changes in the legal context might necessitate either adoption of the proposed tools or application of new technologies. Such should be done, following a conscious evaluation of alternatives and in consultation with other BD4BO / IMI projects and DO→IT.
See appendix 6 for a checklist on data management for reviewers of BD4BO projects to consider.

Appendix A – Points for Reviewers to Consider

This appendix provides a series of points that we recommend reviewers of BD4BO projects consider critically when reviewing proposals to ensure that sufficient attention has been considered by the complexities of data management in consortium working.

- Has the project specified that data will be collected prospectively (Foreground IP) or retrospectively (Background IP)?
- Prospective Data
    - Has the mechanism for collected data from patients in clinic or elsewhere been described?
    - Is their sufficient budget to set up the data processing logistics?
    - Has a technology solution that matches the systems proposed in the document been selected?
    - If no, is a meaningful justification why not given?
- Background Data
    - Is a clear definition in the proposal about how data controllers will be expected to provide data and how others will access the data controller's data?
    - Is there evidence that data controllers are comfortable with mechanism for accessing their data?
        - If IPD will be transferred to a third party is there explicit support that data controllers are committed to the approach (e.g. explicit MoU for each controller)
        - If IPD will be transferred to a third part is there sufficient resource allocated from each data controller to audit the 3rd party for compliance to their institutional standards of privacy, security and ethic and consent governance.
    - Is there sufficient resource allocated to data controllers to prepare, anonymise and transform their data ready for it to be used?
- Technical Platform
    - Has one of the technical platforms highlighted in this document been selected?
    - If not has suitable justification been given why not?
    - Has the justification included as how it will help outcomes research in Europe more generally rather than just in the context of the project?