Data Transparency SIG

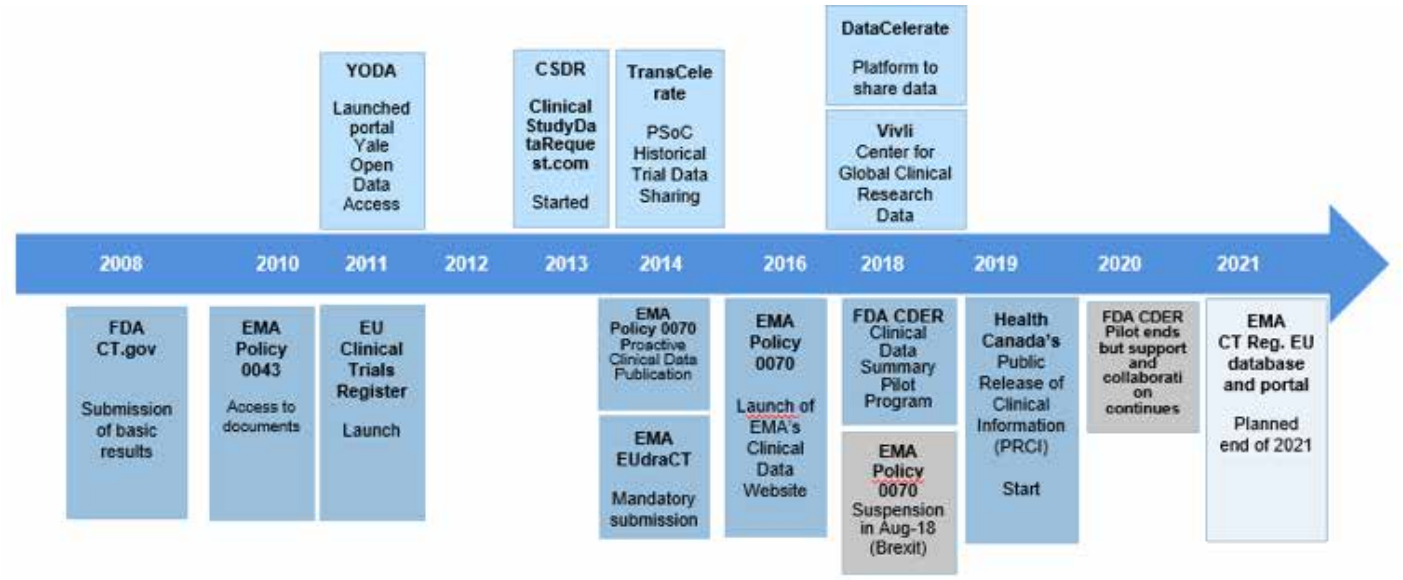
Title: Secondary use of data - Unleashing Data Assets to Create Value

Authors: Janice Branson (Novartis), Mimmi Sundler (Astra Zeneca), Katherine Tucker (Roche).

Introduction

Over the last decade there has been increasing recognition in the value of secondary use of clinical trial data. The data may in fact be valuable for other scientific investigations beyond the initial purposes and objectives of the protocol. Across the Pharma industry and academia, huge amounts of clinical trial data have accumulated over many years. Within Pharma companies as we plan new programmes and investigate emerging and evolving scientific areas, we seek to understand how we can utilise this already collected data. In 2020 with the urgency of a pandemic situation, we also saw a huge interest in collaboration and intentions to share data across companies and academia. What can we learn from the pandemic experience and how can some of these ideas be incorporated into drug development in order to make drug development cycles shorter and more cost effective? **Figure 1** highlights some of the key milestones in the evolution of data sharing and access that we have experienced since 2008.

Figure 1: Key milestones of data sharing and access



There are 4 main areas to consider as we embark on re-use of data: ethical considerations, legal and data privacy considerations, good data science practice and business sensitivity. In terms of ethical considerations, it is important to utilize personal data in line with patients' original expectations e.g. aligned with secondary use language in the informed consent form. However, the best practice for secondary data use may be to utilize anonymized or synthetic data where possible. Data privacy elements are covered in more detail below. In terms of good data science, how do we ensure reproducible research and avoid the reproducibility crisis – can we always ensure analyses plans, data and results follow FAIR (Findable, Accessible, Interoperable and Reproducible) principles? Finally, in terms of business sensitivity, re-use of data should be from locked studies only (or those having reached primary endpoint). Depending on the stage of the program development lifecycle, the project team should be made aware or involved and any commercially confidential information needs to be accounted for. In addition, if the re-use of data is for the actual organization who generated the data in the first place, then some of these considerations may differ slightly compared to external re-use of data. It is true that the external re-use of data has been triggered by public and regulatory pressure on transparency but there is a growing realization that internal re-use of data and the considerations which need to be taken into account are still relatively immature within the industry.

Secondary data re-use programmes internally in an organization

Many larger pharmaceutical companies have multi-year internal data re-use programmes, such as Novartis 'data42', Roche 'Enhanced Data and Insights Sharing' (EDIS) and AstraZeneca internal Data Access Policy (iDAP). The aims of such programmes are to bring early and 'frictionless' access to R&D data across the broader organisation, including making data 'FAIR' (Findable Accessible Interoperable Reusable).

Enhanced Data & Insights Sharing

Clinical trial data has value beyond the initial purposes and objectives of the protocol. Data can be used to inform understanding of patient populations and disease, validate new targets, develop new methodologies, endpoints, biomarkers, tools, and other scientific research thus helping develop more 'personalised healthcare', and ultimately deliver greater benefits to patients.

Data re-use programmes should be underpinned with a comprehensive data governance strategy. Such a strategy ensures responsible data re-use in order to maximize scientific insights, at the same time as minimizing risks, for example those related to data privacy. Responsible re-use could advance science in the interest of patient care, ensure data are used maximally for the public good, bring benefits to society and increase trust in organisations and ethics.

Examples of data re-use objectives include – new insights to characterize the drivers of response to cancer immunotherapy; better understanding the properties of assessment scales used in autism studies (resulting in a change of approach for new studies); enabling clinical and biomarker related questions in a broad breast cancer population, including relationship between certain gene expressions and disease prognosis, thereby informing better designed future studies; identifying groups of super or non-responders.

Data privacy laws and clinical trials

So how do data privacy laws impact our ability to re-use data to answer new scientific questions? 'Data privacy' (US) or 'data protection' laws (EU) cover the fair and proper use of data about people ('personal data' under the EU general Data Protection Regulation or GDPR). Numerous other counties have their own laws, which may diverge from one another and today there are 100+ country specific data privacy laws. So, for a late-stage clinical trial conducted in countries across the globe, this can make a clear assessment of privacy guardrails very challenging.

From UK Information Commissioners Office
What is data protection?

'Data protection is the fair and proper use of information about people. It's part of the fundamental right to privacy – but on a more practical level, it's really about building trust between people and organizations. It's about treating people fairly and openly … and striking a balance with the wider interests of society.'

A pragmatic approach would be to focus on the GDPR as a good starting point. This is true, even with Brexit since the EU currently recognizes the 'adequacy' of UK Privacy laws which aligns with GDPR, at least in the short-term. Under GDPR, clinical trial data is considered as 'personal data' and 'pseudonymised' (i.e. labelled with a pseudonym). All processing of personal data under GDPR comes with a whole host of obligations on the company, institution or other body including a 'legal basis'. In 2019, the EDPB (European Data Protection Board) issued guidance to clarify this for the clinical trials context. It split activities into 'primary use' (and further into 'reliability and safety purposes' and 'research activities') and 'secondary use'. Options for secondary use included analyses deemed 'compatible' to the original trial objective, for scientific research purposes and anonymising data.

Data anonymization

A challenge of anonymisation in the context of global clinical trials are varying data privacy laws and definitions across different jurisdictions. There is no single definition of anonymisation or de-identification (or even terminology). Once data are anonymised, they fall out of the scope of the GDPR (i.e. they are no longer personal data). However, whilst the GDPR does provide a brief definition of anonymization ('with all means reasonably likely to be used, data subjects are no longer identifiable'), in practice it is challenging to interpret what this actually means. The GDPR definition lends itself to a more 'context-driven' approach. That is, considering the overall context and risks of the data sharing scenario as well as the identifiers present within the data.

However, there are now numerous pragmatic frameworks and materials available to work through anonymisation approaches, some developed by industry organisations such as EFSPI/PSI, PhUSE, TransCelerate, guidance as part of more formal routes such as EMA policy 0070, Health Canada PRCI (Public Release of Clinical Information) and publications such as the recently updated UKAN ADF (UK Anonymization Network – Anonymization Decision-making Framework) and those by Prof Khaled El Emam. For example, EFSPI/PSI Data Transparency SIG aim to publish early 2021 on 'Anonymising Clinical Data for Secondary Use' and both Transcelerate and PhUSE are working on common definitions including anonymization. There is also hope that the EFPIA GDPR Code of Conduct on Scientific Research will provide some harmonized industry definition when published later in 2021.

Aside from a choice between personal or anonymised data, a third option to consider is generation of synthetic data. Synthetic data are initially generated from actual personal data but ultimately they do not relate to individuals. However, they retain the same statistical properties of the original data, making them a valuable low-privacy risk alternative for certain secondary use purposes.

Putting it all together: Balancing the demands in an evolving landscape

We have covered the bigger picture of data sharing in the pharma industry as some examples of internal data re-use programmes. We also focused on considerations related to data privacy and anonymisation of data. Aside from these, there are numerous other considerations to take into account when building an internal re-use strategy. Ideally, it should be framed within a comprehensive data and information governance strategy and related policies. Some considerations are outlined in **Table 1**. For example, it is important to consider the flow of data across the entire lifecycle of a clinical trial and when and how in that lifecycle can data be made more broadly available for secondary use (beyond the original objectives of the trial). Development of such a strategy needs to take a truly cross-functional approach to enable detailed evaluation of opportunities and risks.

Table 1: Considerations when developing an internal data re-use strategy

Use original (personal) data	Use anonymized/synthetic data
• The purpose of the re-use is a key consideration • Definitions – primary and secondary use, wider regulatory activities, compatible use, scientific research • Role of ICF and its language regarding data re-use • GDPR legal basis for data processing • Ability to respond to Subject Access Request • Dealing with patient consent withdrawals • How to match planned re-use purpose vs. original protocol objectives and ICF language • Decision-making	• How to define anonymization in an internal context • Definition of acceptable risk threshold • Automation via application of standard rulesets to data • Balancing privacy vs. utility • How to define 'TOMs' (technical and organisational measures) • Complexity across modalities, 'linkability' • Limits on what anon data can be used for (scientific research) • Processes and assumptions for generation of synthetic data
Other considerations as part of a data governance strategy	
• Develop 'secondary use' data privacy strategy – philosophy, principles, definitions, ethical considerations, flow diagram, decision-making etc. • Making data 'FAIR' • Dealing with global studies with patients across multiple countries, implications of different privacy laws • Defining good data science practices – planning and reporting, reproducibility etc. • Defining appropriate internal data access rules • Data flow across study lifecycle, privacy by design	• Systems, processes, infrastructure • Aspects such as reporting/regulatory status e.g. business sensitivity, use of data pre-database lock; safety signal reporting; restrictions due to co-developments, divestments etc. • Other laws and regulations e.g. cybersecurity, use of samples, country specific laws e.g. China 'HGRAC' • Data citizenship & culture, codes of conduct • Staff awareness, training, knowledge • Adapting to changes to external guidance

What can statisticians do to contribute to this discussion? Statisticians are strongly positioned to take a lead in these discussions, both in identifying where secondary use of data may be beneficial within drug development programmes and in ensuring data is collected in a way which will facilitate future data sharing. Statisticians have a deep understanding of the context, trial design and risks associated with a particular trial or molecule and knowledge of the nuances of the data itself. All of these elements are important to inform when data can be made available for secondary use and also in providing this knowledge in the form of associated metadata for the downstream data user.

To conclude, we have seen that external data sharing and collaboration involving the pharmaceutical industry has seen a huge increase in the last 6-7 years via platforms such as CSDR (ClincalStudyDataRequest.com) /Vivli, TransCelerate and others. The COVID-19 pandemic saw a massive influx in discussions across industry on expedited data sharing and dedicated platforms such as Covid19 Data Platform as well as organizations like Vivli and Transcelerate committed platforms to facilitate this. We can learn a lot from the pandemic situation and see how the open and collaborative approach can be applied more generally. It is fair to say many companies have been turning their attention to data as a valuable asset within their own companies and developing internal secondary use policies and programmes. Such a strategy requires a broad cross-functional approach with numerous 'dimensions' to be considered and stakeholders to be included. Such programmes are complex and will develop over many years. Companies need to start thinking about such topics sooner rather than later.

The PSI/EFSPI SIG for Data Transparency considers many of the topics covered in this article: external and internal data sharing and re-use, data anonymization/data privacy in the context of clinical trial data use. If you are interested in finding out more and contributing to the SIG, **please visit the PSI Data Transparency SIG website**, also for some links to the references in this article.