Data Transparency Summer Event 2020 – Presentations and Resources

Welcome to the resources hub of the Data Transparency Summer Meeting. Here you can access the live recordings from the event by clicking below. View each speaker's presentation slides, resources they may have referenced and their Q&A sessions. You can also take a look at what our attendees had to say about our virtual event and industry-leading platform.

Day 1 – Wednesday 3 June:

Speaker	Presentation
Julie Holtzople, *AstraZeneca*	**How Sponsors are Managing Clinical Trial Transparency Requirements Across the Globe -**
Selected Session Q&As	
Are any companies sharing data internally without any kind of anonymisation?	**This is not something we asked in the survey, and it would be a great opportunity for a follow up survey question.**
Are the anonymisation rules different for datasets versus documents?	**They may be, but they can also be treated the same. It depends on a sponsor's process. We did see some variances noted in processes between the two within the survey. We did not specifically seek much insight into rules within this survey.**

Speaker	Presentation
Dr Till Bruckner, *TranspariMED*	**Shortcomings of the Global Trial Registry System During the COVID-19 Pandemic**
Referenced Resources	
Negligence of clinical trial registries undermines COVID-19 medical research	**https://www.transparimed.org/single-post/2020/06/02/covid-ICTRP-clinical-trial-registries**
Results are missing for 1,516 clinical trials of potential COVID-19 drugs	**https://www.transparimed.org/single-post/2020/06/03/COVID-research-waste-publication-bias**

Speaker	Presentation
Alex Hughes, *Roche*	Who is Flying the Plane? A Case Study in Piloting Anonymisation Software
Selected Session Q&As	
What format are the documents, and how do you address the text overflow when replacing 4-digit numbers with 64 digits across the entire document?	The documents are in PDF format. We tested 64-character strings as the most extreme example we could think of requiring as we have experimented with hashing patient numbers. Some vendors do have text reflow capability and where possible we wanted to test the limits of that function while trying to solve a realistic potential problem.
Will the 64-character string be randomised every time, or the same string assigned to the patient ID throughout the documents? How do you then combat linkability?	The same string would be applied to the same patient every time to maintain linkability. For the 64-character strings, for example, these could be generated with a hashing function and a salt value. These one-way functions can help create consistency across data and documents. Another route is creating a "look-up" table and using that to map the desired strings.
From your knowledge, is there any consistent anonymisation of clinical data AND documents ; currently, we do fully parallel desynchronized processes, but is there any requirements /expectations to?	The ultimate aim is to have consistency between the clinical data and the documents (inc. PT numbers, AE /MH generalisations, ages etc...). Our aim is to make the package that we send to the health authorities as useful as possible and to maintain these links. I do not believe this is an explicit requirement (but will happily be corrected) but implicit as redactions/anonymisations have to be justified. Sponsors do have the ability to measure risk, define a risk profile and simulations to use, and then apply them to both the data and the documents, for example having the same scrambled subject IDs between datasets and documents. This option does exist with tools available today.
Why do we only hear about the EMA and Health Canada? What about other parts of the world?	Only the EMA and Health Canada currently proactively make CSRs public.The policies from these health authorities are currently presenting the most immediate challenges for us and many other companies. FDA policies, for example, are not as far reaching when it comes to data sharing. The PMDA does call for publication of some clinical documents, mostly module 2 documents. We are keeping an eye on other health authorities, in Asia-Pacific (e.g. the PMDA, CFDA) and how their policies affect our data sharing approach in the future.
Has the hold on EMA Policy 0070 affected the way they are proactively preparing dossiers for disclosure?	The short answer is no. While the EMA has paused this policy for the migration, Health Canada's policy remains in place and so dominates our current focus. We do not currently proactively treat documents for disclosure, but that is an ambition and the pilot is exploring those possibilities.
Does this software anonymisation for documents happens during the authoring or retrospectively?	It happens retrospectively at the moment. We have thought about anonymisation proactively using tagging, but the current thinking is to adapt the process so that, when the documents are authored, we write them in such a way that lends to being anonymised, i.e. we do not put unnecessary revealing content in.
When converting patient numbers to the hashes, would one patient taking part in multiple studies be given the same hash?	If the studies themselves were linked (by say a SUBJID), then we would aim to keep this link but it really depends on the study set-up. This is more difficult for legacy studies, but if the data and the links are there, then it is possible!
During anonymisation of PDF documents, did you find formatting challenges? Did you resolve these before submission?	Yes! Unfortunately, not all PDFs are created in a standard way. The most common issue I can think of is data that is presented in tables. These things are difficult to fix automatically and often require manual intervention. The challenge is finding all of the cases where this occurred. Another similar issue is text in images.
What were the compromises you had to make with your chosen vendor, e. g. better training package for worse cloud ability?	We are not at this stage yet, but as there is no "push button" solution there will be inevitable compromises. Our priority is quality of output and we will have to balance that with service solutions, features, costs, etc.
Is it cost-effective compared to manual processes?	We don't have a firm answer to this yet. Costs will be considered when making a decision, but it will be more complicated than a simple cost comparison. We will consider other factors, in particular; quality of the output, changes to processes to accommodate the solution, and security.

Day 2 – Thursday 4 June:

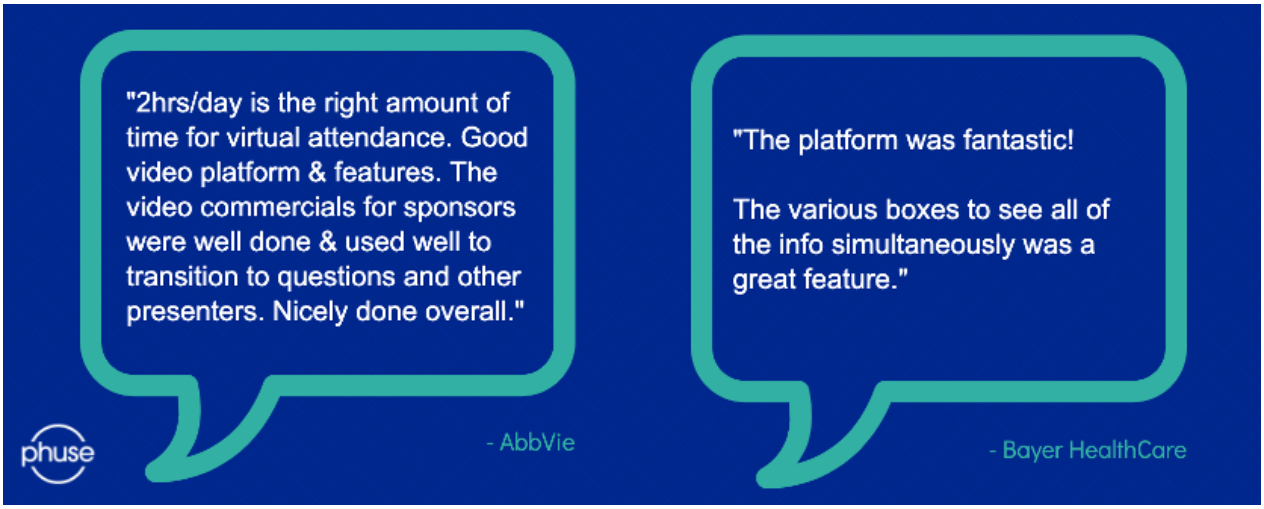
Speaker	Presentation
Luk Arbuckle, *Privacy Analytics*	Hide and Seek: Evaluating Identifiability in an Anonymised Clinical Study Report
Referenced Resources	
Evaluating the re-identification risk of a clinical study report anonymised under EMA Policy 0070 and Health Canada Regulations	https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-020-4120-y
Selected Session Q&As	
Given the description of a 'motivated intruder' is not a specialist hacker but an 'entry level data analyst', does that mean that we don't need to worry about specialist hackers? Especially if docs are in the public domain?	Some of the demonstration attacks I briefly mentioned are examples of expert attacks, mainly academics who can gain notoriety from publication of a novel attack method. Most of these have, however, been on pseudonymised data or data that used ad-hoc methods with no disclosure metrics involved. These attacks are nonetheless factored into the disclosure metrics that exist and how they are used. Reasonableness, however, is judged by the non-expert attacks, which has been borne out in courts and by regulatory review.
Have there been attempts to combine these attacks with red-team/penetration-testing attacks when looking at the re-identification risk?	Red-team would have almost unbounded skills, knowledge, and time. Some of the demonstration attacks I briefly mentioned are examples of those. Mainly academics who can gain notoriety from publication of a novel attack method. Most of these have, however, been on pseudonymized data or data that used ad-hoc methods with no disclosure metrics involved. These attacks are nonetheless factored into the disclosure metrics that exist and how they are used.
What is quantitative anonymisation please? Do you mean quantitative RISK?	The EMA anonymisation guidance recommends a risk-based approach to anonymisation, and allows for two approaches: a quantitative approach and a qualitative approach. The former uses statistical disclosure control techniques to estimate the actual probability of identification. A qualitative approach as has been applied in practice, does not estimate probabilities but uses qualifiers as low/medium/high risk.
Regarding anonymisation and the clustering aspect, which was mentioned, how should the number of individuals in each group or cluster should be defined? Any specific value in guidelines?	The EMA and Health Canada recommend a group size of 11, or 1/11=0.09. This is based on standard practice in statistical disclosure control.
Thank you for the presentation. How would the motivated intruder test work when sharing clinical trial data for research context (e.g: sharing data between two university research groups, who are bounded by data transfer agreement)?	This is the context element briefly mentioned in the presentation. There is a range of practices/controls in statistical disclosure control, to support the use of disclosure metrics in the context of a particular data sharing model. See the '5 Safes' whitepaper at: https://keep-data-safe.com

Were the principles of anonymisation used the same as those used in setting individual datasets to CSR?	The same concepts are used to anonymise SIPD and CSR. The key challenge with anonymising CSRs is extracting identifiable information from documents, running disclosure metrics, transforming that information to render the data nonidentifying, and pushing it back into the document. Natural Language Processing is used, as well as expert review. But after all that, the same concepts are applied.
What are k-anonymity and l-diversity?	You can think of these as clustering methods, they are classified as "similarity metrics". K-anonymity is actually a method first used in statistical disclosure control by the name n-rule or threshold rule. Basically, a minimum count on how many people in a group. l-diversity is a variation on k-anonymity, where "sensitive" or target fields are also considered and evaluated to ensure there's enough variation that the recipient doesn't receive information that is "too specific" about a person. l-diversity isn't really used in practice, because it limits what can be learned from the data even when they can't be identified. I hope this helps. There are actually 80+ metrics in the academic literature, it's a big topic.
Do you think new concepts and techniques (differential privacy) will shortly replace anonymisation?	These newer techniques are disclosure metrics that fall under the umbrella of anonymisation. There are actually 80+ disclosure metrics in the academic literature. Differential privacy falls into a class of "indistinguishability" metrics. It's complicated to explain the differences using a message. I think the methods used in practice have to be pragmatic and scalable, and we will see them borrow from one another to meet regulatory and, importantly, business needs. They all have their pluses and minuses. This is the subject of a long discussion.
Which approach is better, considering reference population - study population or geographical population?	Something in the middle, e.g., the similar trials. Trial population is narrow and will lead to poor data utility. Geographical is too broad, and it will be possible to single out and narrow down to the similar trials.

Speaker	Presentation
Dr. Sarah Nevitt, *University of Liverpool*	Data Requesting and Data Sharing: The Nine-year Academic Experience
Referenced Resources	
Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review	https://www.bmj.com/content/357/bmj.j1390
Antiepileptic drug monotherapy for epilepsy: a network metaanalysis of individual participant data	www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011412.pub3/full
Selected Session Q&As	
Did you work on data anonymised with different approaches for your meta-analysis?	Yes, but this did not impact overall on the analysis of all the results. It just caused some complications for me when preparing data.
How would you handle (is it possible?) a mix between studies anonymised (possibly with different approaches [shift date and relative days] and studies for which informed consent allows secondary use?	The work I was doing really only uses relative days so in theory, as long as anonymisation did not change relative days, I should not have any problem. I would need to carefully inspect the data to check exactly what has been done within anonymisation to make sure of this.
Do you use any software applications for the network diagram generation? Is the methodology to create network diagram published?	I used the 'networkplot' command in Stata.
Have you done any analysis on the time taken by the pharma or academics to provide the data – any comparative analysis?	Yes. Please see the BMJ publication in the resources. We published the time to receiving data.
For this work do you have a break-up of data which were made available to you by various pharma and academic research?	No. Within the network meta-analysis, all studies are included together. I only break up academic / pharma studies when considering factors such as 'availability bias', as I described in the slides. We also published some details about differences between pharma and academic trials in our BMJ paper. Please see the resources.
What more can be done to get academic trialists to share their studies, particularly those who are publicly funded? What are the biggest barriers?	Great question! A lot of progress has been made in the last few years around academic data sharing. For example, we now have Data Sharing policies, procedures and a data sharing team at our clinical trials unit in Liverpool. The UK clinical trials network also has a data sharing subgroup. The biggest barrier will also be resources. We have to find the time and money from somewhere to prepare data for sharing and this is a big challenge for trials which are completed, especially where the original team may be working on other projects. Going forward, we are building in time and funding for data sharing into our research grant requests, to ensure that data sharing will be easier to do in the future.

Do you confirm that the introduction of CDISC data standards during past years has also made your work and analyses much easier?	Yes, absolutely. I certainly found CDISC data models to be helpful as the format that data were provided in was consistent, so it made checking it and preparing outcomes a lot easier. Trials conducted within academic institutions tended to have lots of different formats and did not use recognised data models, so these datasets often required more preparation.
Great presentation. Can you please clarify how the network graph was generated? Using a two-stage or a one one-stage IPD approach?	Thank you. The network plot was generated in Stata (networkplot command). We used a mixed one-stage/two-stage approach to analysis to incorporate the treatment covariate interaction and some additional aggregate data. Please see the published Cochrane Review for full details.
Can you clarify the data format issues? I would think that the CDISC data models would be more helpful now than in the past.	Yes, I certainly found CDISC data models to be helpful as the format that data were provided in was consistent, which made checking it and preparing outcomes a lot easier. Trials conducted within academic institutions tended to have lots of different formats and did not use recognised data models, so these datasets often required more preparation.
What are the main challenges you experienced in requesting/receiving the data from the sponsors?	Previously (pre-2014), it was actually making contact with sponsors, but that is no longer a problem. Now, it is very straightforward to request pharmaceutical data and to have ongoing conversations about the data if I have any questions about it. Remote access to data remains a challenge as the methods I use require me to have all of the data in one place. I am aware that some companies allow download of data under certain circumstances. I am interested to see how allowance for combining data across different platforms develops.

Speaker	Presentation
Prof. Khaled El Emam *Professor, University of Ottawa & CHEO Research Institute & Director, Replica Analytics*	Experiences with Synthetic Clinical Trial Data
Referenced Resources	
Data Synthesis Tutorials	https://replica-analytics.com/synthesis-tutorials
Data Synthesis Book	https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/



Day 3 – Friday 5 June:

Speaker	Presentation
Prof. Joe Ross, *Yale University*	The Value of Data Sharing: Lessons Learned Through the YODA Project
Referenced Resources	
Overview and experience of the YODA Project with clinical trial data sharing after 5 years	https://www.nature.com/articles/sdata2018268
Selected Session Q&As	
Through the YODA Project, would we get access to collected data or analysis data?	Both CSRs and raw IPD (the analysis-ready datasets) are made available via the YODA Project.

Speaker	Presentation
Dr Jennifer E Miller, *Bioethics International*	Recent Feedback from the Good Pharma Scorecard
Referenced Resources	
White Paper	Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices
Selected Session Q&As	
Does your NDA and patient trials reporting include trials that have not reached their primary completion date over one year? Sometimes a product may be approved, but studies included in the NDA may not have completed.	Ongoing trials at time of approval are excluded (by PCD).
Can you just clarify the difference between the 'All NDA trials' slide and 'All patient trials' – is the difference healthy volunteer trials? Or something else?	The patient trial sample generally includes (1) completed trials at time of approval (2) conducted in patients (3) for the approved indication. It excludes trials for unapproved indications. Phase 1 trials in healthy volunteers, observational trials, expanded access and a few other conditions from results reporting.

Speaker	Presentation
Benjamin T. Rotz, *Eli Lilly & Co*	Clinical Trial Transparency: An Industry Perspective of What Was, What Is, and What May Be

Speaker	Presentation
Prof. Frank Rockhold, *Duke Clinical Research Institute*	Open Science, Data Sharing and Reproducibility: What's all the Fuss About and Why Should we Engage in the Journey?
Referenced Resources	
Reference Paper	Incentives for Clinical Trialists to Share Data
Selected Session Q&As	
Do regulatory agencies like HC and the EMA take a data controller role? Or is it a shared responsibility with sponsors, or solely of sponsors?	I guess it depends on how you define the role but I guess I would say neither – an independent review panel is the best way to implement that.

Fridays Panel Discussion:

Watch the Panel Discussion from Fridays recording for insight into this conversation, and see what the panel answered to the below questions.

Panel Discussion Questions
Can you share your thoughts and guidance on how sponsors can improve internal processes for reviewing, approving and preparing data to be shared from a data sharing request from a resourcing, time and cost perspective?
In what different ways may patients may be involved with disclosure of data?
Recent events in the US have exposed the gaps that exist in medical care across communities in this country. Would Dr Ross and the panel comment on the possible role data sharing may have in helping to address this gap, e.g. sharing of data for meta-analyses of diabetes and CV trials?
When is the next GPS going to be published? Will companies be informed and have a chance to respond, as they did last year?
Im interested to hear how disparate public document policies across the EMA, HC, the FDA and others could be aligned.
Is there any guidance as to whether or not to de-identify control arm data if it is proposed to be analysed and included as evidence to regulatory agencies?
Re-cost of data sharing should be sustainable. How do you think it should be funded?
Looking five years on, what changes do you expect to see in data sharing?
How strongly does the panel feel that a secure data access system must be used? The down side is that it hinders data utility.

Joe pointed out how sharing platforms aren't working together. What are the reasons for this?