UK Health Data Research Alliance

# Pan-UK Data Governance Steering Group

## 09 October 2023

**Chair:** Andy Boyd

## Summary of Key Discussion Points

- Andy Boyd welcomed delegates and introduced the meeting.

## Synthetic Data

***Are Synthetic Health Data 'Personal Data'?* – A report by Colin Mitchell and Elizabeth Redrup Hill, commissioned by CPRD and led by Dr Puja Myles (found here).**

- Elizabeth Redrup Hill introduced the report. Summary of main findings:
- The CPRD definition of Synthetic Data was adopted in the report. The report looked at the circumstances where synthetic data can be considered patient data. Synthetic data is created using real data, using both manual and automated methods, and closely resembles real patient data. The risk of identification is on a spectrum and includes nature, environment, context, and technical and organisational safeguards during anonymisation. A key part of assessing reidentification risk is looking at how sensitive the real data is.
- Therefore, there are privacy risks which cannot be guaranteed to be removed during the process.
- A holistic approach to evaluating the patient privacy risks associated with synthetic data is necessary – different methods are required depending on the type of synthetic data and its uses. A risk/ reidentification assessment should be done on all synthetic data, including on where this is stored and how likely reidentification is.
- An orthodox approach could lead to slow data availability, risk-averse conclusions, and uncertainty in how to apply data protection legislation and obligations.
- If we treat all synthetic data as personal data, we might overstretch legal certainty. An alternative approach is to presume that some synthetic data is not personal data but to add additional safeguards to prevent reidentification, such as removing outliers.
- A consensus must be reached for clarity on the standards and approaches to identifiability – a proportionate response should be adopted where reidentification risks are low.

**Discussion points:**

- Public mistrust of synthetic data – it's a complicated subject and can seem frightening, as synthetic data is a wide term it has many different interpretations and there is a challenge in communicating this to the public. Work has been done to gain public trust, such as motivated intruder tests and removing outliers and accidental matches, however more work is needed in this area.
- There are issues with research using synthetic data being published as "real data" – it would be useful to watermark this data to ensure that it is not being used for research.

**Case Study - *Real World Applications of Synthetic Data – leveraging national cancer data to accelerate insights for patient benefit.* Julia Levy, Oncology Real World Studies Principal, UK and Ireland, Lora Frayling, Health Data Insight CIC Partnership**

- Determining the value of synthetic data starts with recognising the value of insights that real-world data can bring to patients and the clinical community. In the field of cancer research, data is a critical resource in decision making and in the lifecycle of drug development.
- In England, Cancer data is held on England's Cancer Analysis System (CAS) which brings together multiple datasets, including data on cancer diagnosis, molecular testing in routine clinical practice, admissions and hospital episodes, chemotherapy, and ONS data.
- Understanding the full patient pathway can help researchers establish e.g. treatment patterns and clinical needs and optimise clinical care.
- Though the CAS is incredibly useful and impactful, it is very difficult to access directly from NHSE and the process of obtaining access is time consuming. This is less than optimal for patient care.
- To accelerate the research process, Simulacrum -synthetic data created from the CAS – was created using real data with the relevant properties but no real patient information.
- Simulacrum allows researchers to efficiently create code to navigate the CAS without having to ask for access to the actual data repeatedly, significantly speeding up the process to generate insights.
- Simulacrum was created through a partnership with IQVIA,HDI and AZ.
- Simulacrum is created by using an algorithm that transforms the data into anonymous aggregate data tables from which synthetic data are then generated. Since the synthetic data are derived from anonymous data, i.e., not personal data, we can easily answer the question "is this synthetic data considered personal data?" as "no", since it is not derived from personal data.
- Risk of patient reidentification is assessed at each stage of the generation process: i.e., risks from (1) the real data, (2) the fitted model (i.e., aggregate tables) and (3) the synthetic data.
- Risk from the real data is assessed through evaluation of data variables, followed by selection and pre-processing to reduce their sensitivity.
- Risk from the fitted model, i.e., the anonymous aggregate data tables, are assessed according to NHSE data release guidelines. The method to fit the model uses a clustering algorithm that ensures they fulfil requirements for release and are therefore considered anonymous and releasable.
- Risk from the synthetic data are assessed through post-hoc evaluation. This includes metrics seen in most current literature, including the replication of outliers. No two identical records exist in Simulacrum when looking at the full patient record.
- It is important to keep track of all released outputs derived from a single data source, including routine data releases, synthetic data and information about its evaluation and synthesis methods, as these can all impact data privacy.

**Discussion points:**

- The perception that someone has found something can be just as harmful to a data owner as whether they have. In America, social scientists have invested in replication services to run analyses between true and synthetic data. This may be helpful.
- Misconceptions around synthetic data undermine patient trust that health data is not being distributed. It might be better to have more secure environments to work with real data.
- Creating the anonymous data tables in the Simulacrum model involves merging patient groups with rare combinations of variables with similar groups to ensure sufficient anonymity. However, this can sometimes cause loss of detail and bias in the generated synthetic data., i.e., lower fidelity. There is a trade-off between privacy and fidelity. However, this does not impact utility in IQVIA's use case as it does not affect final analysis results from the real data.

## Action force updates:

- The Pan-UK Data Governance Steering Group is writing a consensus/call-to-action paper for integrated and harmonised data governance across the UK. The paper has been through one round of reviews and comments are being worked on. This will then be sent to members of the group to review and contribute.
- **TRE Legal Toolkit Action Force**: We are currently looking to rationalise and standardise legal contracts for data access in TREs, working with the UK TRE/SDE network. We are very keen for ongoing conversations on this, especially from those who wish to deploy the toolkit in future. The first working group paper has been published in the IJPDS Special Issue: Advances and Innovation in Data Governance and the Data Access Agreement has been released for adoption.
- **International Data Access Action Force**: The first meeting was held where a group of organisations looked at the challenges of international data access and the ability to provide a service where data is shared from a secure UK location, and what the legal and auditing challenges would be. HDR UK has taken an action to map out different issues and will present this to the group.
- **Transparency Standards Call for Funding**: We received 24 applications which are currently under review, with decisions to be made in October. The funding is to support organisations to meet the transparency standards, with outputs to be delivered by March 2024. The number of applications illustrates the need for funding in this area.
- **Issues in the news**:
  - UK- US Data Bridge;
  - The NHS issued a call on public engagement and communications on Health Data;
  - NHS SDE programme update on the Data Access Policy.

## PPIE Updates:

Ester Bellavia gave an update on the work of the PPIE team.

- **The PPIE Strategy and Steering Group** has met once and is meeting again to focus on four areas;
    - A best practice approach to high-quality meaningful PPIE
    - UK Health Data Research Patient and Public Collaboration Forum
    - Activities that build awareness, understanding, and support of health data research across the UK population.
    - Involving frontline healthcare professionals.
- **Humber Science Festival**: The HDR UK PPIE team participated in this event, with the target group being young people and families. The event was a big success and there are plans to attend more of these types of events in future. HDR UK plan to attend at least one festival in each of the four nations within the next year.
- **HDR UK Voices**: This campaign has been advertised through different channels with several contributors involved in this since it started. It's been highly successful so far. See here.
- **PEDRI**: PEDRI introduced seven best practice standards last year (Equality, Diversity, and Inclusion; Effective Communication; Data Literacy and Training; Proactive Transparency; Mutual Benefit; Meaningful Involvement and Engagement; Culture of PIE).

## Four Nations Updates:

**Wales - SAIL Databank:**

- The Welsh Government has developed data for research including SAIL and the development of policy to find, recruit and follow up data participants and organisations. The Welsh government has been collaborating with SAIL to develop a business case to evolve this process. There is a second element of data led clinical trial development; a series of projects are on the go and the HRA is leading a project on data consent. A tender is out to set up a mechanism to contact potential participants. Lastly, the Welsh Government is working with NIHR to work out how best to use the volunteer register in Wales.
- Within SAIL, the Information Governance team are extremely busy; applications were at an exceptional level throughout the pandemic but are still very high. The SAIL team are very grateful for the production of the transparency standards and have applied for some funding to create a video. SAIL adheres to a great number of the standards already but are looking to address any gaps. SAIL is reviewing the template DAA created by the TRE Legal Toolkit Action Force for adoption. SAIL is very engaged with the concept of synthetic data and is looking for some grant funding options for projects, though there are some issues with governance pertaining to anonymisation. The synthetic data would be used for student researchers mainly as a bridge to access more complex datasets.
- An article around data federation is in the process of being published.
- There is an annual review of accreditations and certifications within SAIL as the ISO audit is due this year. SAIL is developing a trust programme and welcome suggestions on transparency when publishing audits. SAIL aim to improve confidence in self-reporting.
- SAIL is actively involved in discussions around international data transfers. There is engagement with IT and local site security to enhance security, this is a complex process and advice is welcome especially on working with international audiences.

**Northern Ireland – Honest Broker Service**

- The Honest Broker Service is going through some updates to streamline approvals. There are ongoing issues with secondary use legislation as this work was not completed before the introduction of the Digital Economy Act, so the work is ongoing with the Lead of the Department of Health but there may be a need for new legislation. It's hard to predict when these issues will be ironed out.
- A new computer system has been implemented, where all data will be hosted covering several data source areas. At the moment, data availability differs depending on the Trust. There will be a transition phase for ingress of data, including setting data standards and managing the process with researchers.

**England – NHS England Data for R&D Programme**

- The Data Access Policy is currently being considered by the Department of Health and Social Care; there will be an update on this coming out shortly and the update will be shared with the group. Many have responded to the consultation put out around policy and the direction that it is taking.
- The SDE network aims to implement data access across the NHS, there are eleven subnational data access environments and one national one in NHS England. The information governance is currently being considered. A vision has been mapped in which there is one gateway to access all SDEs. NHSE are working with HDR UK closely to work on developing a single front door that leads to all SDES and aggregates the data available so that users can select what they wish to access. There is a single form being developed, which is proving to be a challenge. This model would be a federation model which means that there will be twelve data controllers, with the potential to increase, and the end user still feels like they are accessing a single place.
- There are various workstreams that the information governance community are working on. The key challenge is in finding consistency and commonality in processes for the sub-national SDEs to work to.
- There has been adoption of standardisation wherever possible through series of self-assessment resources available to all SDEs.

**Scotland –Research Data Scotland**

- Research Data Scotland (RDS) has been continuing to work with the Scottish Government and other partners and have been involved in conversations regarding information governance code of conduct.
- RDS are working with DARE UK on projects (SACRO consensus statement, SATRE principles)
- 'Scotland Talks Data' – a refreshed public engagement panel.
- Developments on the RDS website and Researcher Access Service, with prototype application forms user tested and video overviews shared with RDS partners.
- Continued work with ADR Scotland and support of the Scottish Safe Havens.
- Work with partners to develop policy around industry access to public sector data.

- Scoping the synthetic data landscape and strategy, having set up a Scottish working group and joined the UK synthetic data group. RDS will take over Synthpop management and have funding to investigate synthetic data tolls, provide guidance to data controllers on disclosure risks and develop example synthetic datasets.
- Launch of Synthetic Data Fund in October 223 as part of RDS Systems Development Fund. Estimated allocation of £100k for synthetic data research funding applications.

## Open Forum:

- Information security and compliance- SAIL are working on review of policy and procedures to streamline and reduce the burden of audits. SAIL is looking to collaborate with other members of the Pan-UK Steering Group as there is a lot of scope for getting it right.
- The future is not evenly distributed and there is a need to support this by sharing best practice and use tools available to prevent duplication. There are opportunities for partnership working across the SDE network.

## Attendees

| Name | Organisation |
|---|---|
| Janet Valentine | ABPI |
| Emma Gordon | Administrative Data Research UK (ADR UK) |
| Neena Modi | British Medical Association (BMA) |
| John Lathan-Mollart<br>Puja Myles | Clinical Practice Research Datalink (CPRD) |
| Hans-Erik Aronson | Data and Analytics Research Environments UK (DARE) |
| Tim Hubbard | Genomics England |
| Cassie Smith<br>Edel McNamara<br>Rachel Brophy<br>Eilidh Ferguson<br>Ester Bellavia | HDR UK |
| Lora Frayling | Health Data Insight CIC |
| Paola Quattroni<br>Uwaye Ideh | Health Data Research Alliance, HDR UK |
| Scott Mathieson | Health Social Care Northern Ireland |
| Julia Levy | IQVIA |
| Vicky Chicco | National Data Guardian Office (NDG) |
| Gary Coleman | NHS E |
| Claire Edgeworth | NHS R&D |
| Gary Ricker<br>Rebecca Jones | Office Life Sciences |

| | |
|---|---|
| Bill South | ONS |
| Maeve Groot Bluemink | Our Future Health |
| Elizabeth Redrup Hill<br>Peter Mills | PHG Foundation |
| Sara-Jane McAteer<br>Munisa Hashimi | Public Advisory Board |
| Kate McBay | Research Data Scotland |
| Sharon Heys | Secure Anonymised Information Linkage Databank (SAIL) |
| Andy Boyd | UK Longitudinal Linkage Collaboration (UK LLC) |
| Felix Ritchie | UWE Bristol |