PhenoMeNal

Deliverable 7.1.2

Project ID	654241
Project Title	A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data.
Project Acronym	PhenoMeNal
Start Date of the Project	1st September 2015
Duration of the Project	36 Months
Work Package Number	7
Work Package Title	Privacy and Ethics
Deliverable Title	D7.1.2 Workshop 3 on best practices in handling sensitive human data, taking into account National and Institutional legal policies
Delivery Date	M30
Work Package leader	ICL
Contributing Partners	ICL
Authors	Robert Glen, Timothy Ebbels, Noureddin Sadawi
Abstract	A workshop was held on 21st February 2018[1] to further discuss the key role of Ethical Legal and Social Implications (ELSI) within the phenomenal project. The main task was to examine the impact in two areas: the new GDPR legislation and the technical challenges in making PhenoMeNal safe and secure. Four speakers from different backgrounds (law, security, software and clinical) gave lectures outlining the challenges and solutions in their areas of expertise. Outputs from the workshop include key learnings on handling sensitive

[1] http://phenomenal-h2020.eu/home/2018/02/21/elsi-workshop/

	data (technical and procedural) and tasks to maintain PhenoMeNal GDPR compliant.

Table of Contents

1 Executive Summary

A workshop was held on 21st February 2018[2] to further discuss the key role of Ethical Legal and Social Implications (ELSI) within the phenomenal project. The workshop follows on from two previous Phenomenal workshops on ELSI. The main task was to examine the impact in two areas: the new GDPR legislation and the technical challenges in making Phenomenal safe and secure. Four speakers from different backgrounds (law, security, software and clinical) gave lectures outlining the challenges and solutions in their areas of expertise and the relevance to Phenomenal. As we as informing and educating the group in these areas, a number of key tasks and actions were identified.

2 Contribution to the project objectives

This deliverable has contributed towards the following project objectives:

- Raise awareness of information governance within the consortium and assure ongoing compliance
- Provide a forum for information exchange on best practice in clinical data sharing and disclosure
- Ensure that ethical standards and guidelines of Horizon2020 will be rigorously applied regardless of the country in which the research is carried out.

3 Detailed report on the deliverable

Privacy and Ethics continues to be an important consideration in both the software design and conditions of use in Phenomenal. A third ELSI workshop was held during this period to finalise the discussions on this area within the project.

3.1 Workshop Key Objectives

- To understand the developing environment for ELSI issues governing patient data.
- Security of the PhenoMeNal Infrastructure including General Data Protection Regulation (GDPR) implications

3.2 Report on the workshop

An introduction to the ELSI area and Phenomenal was given by Robert Glen. He outlined that the main focus of this workshop was the impact of GDPR and also

[2] http://phenomenal-h2020.eu/home/2018/02/21/elsi-workshop/

technical challenges in implementing the Phenomenal infrastructure to comply with current and future legislation and provide a sustainable architecture.

3.2.1 Gauthier Chassang

Presented a talk on how 'Prepare UU GDPR (General Data Protection Regulations) compliance in e-infrastructure processing sensitive personal data'. This covered the essential legal definitions of the GDPR, the new approach being taken, the role of the Data Controller and the Data Processor, Individual rights and sanctions if the process was not followed. The definition of which data was personal, refers to an identified or identifiable natural person (the data subject) and includes pseudanonymised data but excludes fully and irreversibly anonymous data. Sensitive data in our area includes genetic, biometric, data concerning health etc. The specific processing prohibited by principle and limited exceptions were covered. More definition was given to data processing, which concerns us in phenomenal. Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, for an identified purpose. The legal entities concerned for our purposes are the Data subject, the data controllers, the data processors, the data protection officers (DPO), and the supervisory authorities. The new approach to GDPR is based on potential risks, accountability, transparency, updated general data protection principles and recognised ethical standards. A long-term approach from data collection to data destruction is covered. There is also an extraterritorial scope, outside the EU. The new obligations under GDPR include taking into account the nature scope and depth of the data, new technologies, the nature of the data, large scale processing, vulnerability of subjects and the likeliness of risk. The data protection officer is a mandatory position. A number of key tasks were outlined. Inform and advise the controller or the processor and the employees who carry out processing of their obligations; monitor compliance of policies, contracts, including in the assignment of responsibilities; awareness-raising and training of staff involved; advice in the performance of DPIA; cooperate with the supervisory authority; maintain updated records of data processing operations. A data protection impact assessment should be prepared. There are guidelines and tools for this, ICO (Information Commissioners Office), CNIL (Commission Nationale de l'Informatique et des Libertés) etc. and European Guidelines Art.29 DPWP WP 248rev.01, 4 October 2017. The rules to be followed after a data breach were gone into detail. Notification is mandatory. We should proceed as follows. On Suspicion of a personal data breach - immediate initial investigations. Confirmation of a breach - immediate notification of the data controller. As soon as the controller is aware of the breach, the DPO must be involved and a 72 hours delay starts for notifying the supervisory authority. Nature of the Confidentiality breach and the Integrity of the breech breach as well as the availability of the breach should be evaluated. An Analysis of the breach causes, pathways and impact or risks of impact (collaboration is crucial) is required. At the same time, the controller and processor must act to contain and recover the breach.

A flowchart was given to show the process of handling a breech. He then described problems with imperfect harmonisation in the EU, and the updated rights to information and transparency. The updated notions on consent were important.

These are: Opt-in only, Given for one or several purposes, with granularity (choice/options) for the data subject. In research, data subjects should be allowed to give their consent to certain areas of research or parts of research projects to the extent allowed by the intended purpose and ethical standards. E-consent / Oral consent are possible if allowed by National laws, distinct from any other engagement and consent withdrawal shall be ensured and made as easy to give than consent. The re-use of scientific data was examined. An important point is that processing should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separates from that which allowed the collection of the personal data is required. The new rights to be 'forgotten' and the right to data portability were discussed. Sanctions were discussed. Up to 20M euros for public organisations and Up to 4% of the total world-wide annual turnover for private entities as well as penal sanctions regulated by Member States.

Points discussed with the audience:

- Whether GDPR provides a clear definition of anonymisation and anonymised data as this can be a difficult area
- The new rights to be 'forgotten' and the right to data portability were discussed.
- Whether Phenomenal is a Data Processor, this depends on the data being processed. Should always be anonymised if not in a controlled environment.
- Sanctions were discussed and who is responsible within Phenomenal.

3.2.2 Ben Glampson

Described the work going on at Imperial College in the National Health Service (NHS) research informatics program the NIHR (National Institute for Health Research) informatics collaborative and electronic patient consent for research and clinical trials. The program aims to develop information governance and de-identification procedures to enable access to routinely captured clinical data, connect and warehouse data from many clinical systems, make data and metadata available to researchers, to Introduce new systems for processing structured and unstructured data and to translate research into clinical improvements. He described the information governance challenges. There is also the large problem of capturing unstructured clinical data (7000 employees), in a paper-based culture. He outlined the progress from a fairly small implementation rapidly to a much larger more ambitious project linking several hospitals and therapeutic areas to clinical researchers. He showed a roadmap of future developments up to 2019. This was an integration of Structured data with unstructured data, GIS integration, fully

catalogued data warehouse, pipelines to HPC (High Performance Computing Group) and DSI (Data Science Institute), Cohort identification based on combined clinical ontologies, full Integration of clinical trials into EPR (electronic patient records) , HIEDW – Cerner health-e-intent (Cerner is a company that provides an electronic records system) and a clinical returns program is allied to this, and concentrates on large patient databases in therapeutic areas such as cardiovascular, viral hepatitis, critical care and renal problems. The critical care cohort has 30,000 admissions with 85M individual observations. He then observed on the importance of metadata and how they were treating that separately from clinical datasets.He also talked about the move towards more universal patient consent for research purposes and bio-banking. This included a consent portal and an electronic consent app.

Points discussed with the audience:

- The importance of metadata - and the speaker explained how they treat it separately from clinical data
- A discussion on developing consent forms took place with the speaker. He talked about the move towards more universal patient consent for research purposes and bio-banking.

3.2.3 Dylan Spalding

Talked about the European Genome-Phenome Archive (EGA). The overall architecture, submission and distribution of data, developments and accessibility and security of the EGA within GDPR. The EGA is a joint service from the European Bioinformatics Institute (EMBL-EBI), UK and the Centre for Genomic Regulation (CRG), Spain. It is a secure and permanent archive for all type of genetic, -omics and phenotypic data from humans and ALL data is consented for research use. Data access is managed through application and encrypted data delivery and submissions include raw data from genome sequence, transcriptome, epigenome experiments, called variants and genotypes and sample phenotypes. There are over 10000 data access accounts, 650 submission accounts, 5PB available for download, 3800 datasets and over 500 data access committees (!). It covers a number of diseases from cancer to inflammation. There is significant growth each year. Submitters and users are all over the world. People submit to EGA to maximise the use of their data and because funders and journals mandate this. He then described the EGA Criptor, a system for encryption prior to distribution of data. Metadata is also included to describe the project, samples, experiments and analysis. EGA also asks for the data access committee for each dataset, the policy on terms of use of the data and the data is put into package files for distribution. Access is controlled by the data access committees. The archival process at EGA involves logging access, a private key is needed to decript/encript and an archive key and EGA private key needed to archive files, hosted on shared EBI storage. He showed the overall data access mechanism and the data access agreement for one project as an example. There are over 500 DACs (Data Access Committees) with each having their own access agreement. He

described their secure streaming service for data distribution. EGA has multiple collaborators, but 2 of the main ones we are working with to improve the access to EGA data are ELIXIR and GA4GH. He gave an example case of a query for a variant in EGA (a polymorphism in DNA sequence). He described Beacon, 3-tier access to public, registered controlled data. Each Beacon implementation can implement its own access policy. The process of authentication and authorization is in collaboration with ELIXIR. This allows a user to get an ELIXIR identity, and associate it with a group, role, or institution, so the user could associate their identity with a home institution and a consortium for example. The identity can be linked to ORCID or social media. It also allows for step-up authentication which is a requirement for EGA DACs. It allows for a user to be identified as a 'bona fide researcher', which could be defined as a researcher having a publication in PubMed for example. This 'bona-fide status' could be used for registered level access to Beacons for example. He then described consent codes and how these are implemented. This groups datasets together in terms of consent agreements. Researchers can search by access requirements which clarifies the access requirements. The consent requirements machine readable and are mapped to an ontology (DUO – Data Use Ontology) which will help to automate dataset application / access. He described the GA4GH (Global Alliance for Genomics and Health) streaming API which allows access to allowed data as if it were open access. Other data access approaches are the Resource Entitlement Management System (REMS) which is being developed by CSC (which is an organisation owned by the state of Finland and the Finnish Higher Education Authority) and which we are trying to integrate with the EGA. Secure cloud access using a federated data system, e.g. for UK-Biobank, an advantage of the cloud is the number of users using a single copy of the file, unlike UK-Biobank for example (multiple copies of very large datasets). The requester also gains access to data within a secure cloud and can perform their analysis, without having to directly access the EGA at all. They continue to review security within the GA4GH security working group and do regular risk assessment and risk mitigation. He also described how GDPR will affect their efforts, not much as they are working towards a GDPR compliant environment at EBI.

Points discussed with the audience:

- Advantages of the cloud were discussed such as the fact that the number of users using a single copy of the file can be large, unlike the way it's implemented in UKBioBank (multiple copies of very large datasets)

3.2.4 Kenneth Haug

Discussed security, architecture and designing with ELSI and GDPR within PhenoMeNal. The technical solutions implemented in Phenomenal introduce a number of challenges. To ameliorate these, there is Network and physical access control (not ELSI):

- Only allow secure web access

    - SSH, SFTP, SSL/HTTPS
    - Access using private/public keys, not passwords
    - Firewalls stopping access and unwanted protocols
    - Two-factor authentication
- Container sources
- Container runtime (isolation)
- OS and other software updates (cluster awareness)
- 3rd party (workflow) software “leaking” information
- Code injection
- Inference and guessing passwords

We have discussed Federation, and during the meeting this was examined more from the security point of view, which again raises many challenges. It is not only scientific data that is under GDPR, but also data collected during registration, geographical and administrative data. Examples include:

- User accounts, Single Sign-On (SSO)
    - Elixir accounts, Google, LinkedIn, ORCID, EduGain
- Galaxy and Jupyter accounts in PhenoMeNal
- Google Suite (Drive, Hangout and email)
- GitHub, Slack, Pivotal Tracker, Support tickets, WordPress, CI, Zenodo, YouTube
- Email distribution lists
- Linux (OS accounts)
- Log files (local and external), IPs, MAC addresses etc

ELSI introduces architectural constraints on how the data can be obtained, stored and used. This includes:

- the study design and consent
    - retrospective consent is expensive and difficult
- Moving or sharing data
    - Institutional, National, International
- Local disk vs external clouds
    - No consent = behind local firewall(s)
    - Consent = EGA model for privacy-controlled data
    - Public data = Anonymised data in public repositories (MetaboLights)
- Data retention policy
- Data Use Ontology
    - Identify data restrictions per data set
    - Compatible with institutional data use policies?

For EMBL-EBI, there are a number of questions. Is it an insider or outsider (in EU), and how does it factor in international collaborations outside of the EU (EMBL expect to have outsider status from May 2018). There are restrictions on what insiders can transfer to outsiders. Another point is the right to erasure. EMBL is not obliged to

erase personal data, since this is carried out for the public interest, and it is responsible for the protection of the data (as a whole) (that is the EMBL position). New privacy and consent forms are under development taking GDPR into account. We discussed cloud deployment, and how this is impacted by data federation, security, geographical location, conditions of use.

Points discussed with the audience:

- Data federation and whether we make sure people agree on what it means and how it is explored in PhenoMeNal
- The status of the EBI with regards to GDPR. Is it an insider or outsider (in EU), and how does it factor in international collaborations outside of the EU (EMBL expect to have outsider status from May 2018).

3.3 Summary

This was a wide-ranging discussion covering all the issues of ELSI and technical security in Phenomenal. The speakers highlighted the relevance of each of the key issues in respect of GDPR and in consent/access and security, of relevance to Phenomenal. There are a number of summary points that we could take from these detailed discussions:

- GDPR regulations should be both understood and acted on where appropriate, with advice from more experienced groups such as GA4GS, EMBL-EBI. There are many complexities, national differences and legal technicalities.
- There is considerable flexibility in data use for societal benefit, especially for clinical benefit. This allows flexibility in GDPR for appropriate cases.
- It is important to get local ethical approval when conditions change from those initially envisaged when consent was given.
- We should have plan for a breach, as a potential data processor (irrespective of the fact users should use only anonymised data).
- The data protection officer is a mandatory position. A number of key tasks were outlined. We should understand these. This has implications for sustainability of Phenomenal after the end of the project. Who will be responsible?
- There is considerable flexibility in data use for societal benefit, especially for clinical benefit. This allows flexibility in GDPR for appropriate cases.
- It is important to get local ethical approval when conditions change from those initially envisaged when consent was given.
- We should have plan for a breach, as a potential data processor (irrespective of the fact users should use only anonymised data)
- The data protection officer is a mandatory position. A number of key tasks were outlined. We should understand these. This has implications for

sustainability of Phenomenal after the end of the project. Who will be responsible?

- Much progress seems to being made ‘behind the firewall’ in Hospital/health care environments. How do we ensure Phenomenal is integrated into this process (Ben mentioned they did not have any call for metabolomics, as yet).
- EGA has a gold-plated security solution, how can we take advantage of that, in collaboration, to improve access to metabolomics data processing? We should discuss, e.g. in the area of data federation.

4 Delivery and schedule

This task was delivered on schedule.

5 Conclusion

ELSI continues to be a vital component of delivering a sustainable infrastructure in metabolomics. The developing environment of GDPR legislation and significant challenges in making data secure, available and safe in an ELSI compliant environment will of necessity be a key component of future iterations of the Phenomenal architecture. We will continue to engage with experts and opinion formers in the area to keep abreast of developments and implement these as required in Phenomenal, to maintain its usability within the community.

6 Annex

6.1 Workshop Agenda

13:00 - Introduction. ELSI landscape and Phenomenal. Robert Glen

13:15 - Legal and social issues for handling clinical/scientific data. (30 mins presentation, 15 mins discussion). Gauthier Chassang (INSERM)

14:00 - Combining patient data with scientific research - warehousing, data-sharing, consent and strategy. (30 mins presentation, 15 mins discussion). Ben Glampson (Imperial College London)

14.45 - The European Genome-phenome Archive: security and ELSI update. (30 mins presentation, 15 mins discussion). Dylan Spalding (EMBL-EBI)

15:30 - Tea/Coffee break

16:00 - PhenoMeNal: security, architecture and designing in ELSI. Kenneth Haug (EMBL-EBI)

16:30 - Summary and Discussion.

6.2 Speakers

Gauthier Chassang is a lawyer in International and European law specialised in the field of scientific research involving human beings and the use of biological samples. Working at the INSERM UMR1027/US13 he is part of the French National Node of BBMRI-ERIC, the BIOBANQUES Infrastructure established in Paris. His main fields of expertise and counselling relates to human rights protection in scientific research, such as privacy and data protection in the development of biobanking and new health-related technologies such as whole genome sequencing, synthetic biology and e-health.

Ben Glampson is NIHR-HIC programme manager at Imperial College Healthcare NHS Trust. The National Institute for Health Research Health Informatics Collaborative (NIHR HIC) is working to make anonymised NHS clinical data more readily available to researchers. This involves dealing with the complexities of the legal and ethical environment that are often barriers to research. His team developed systems and software to allow secure access to researchers utilising a wide range of clinical data both within and outside of the National Health Service in the UK.

Dylan Spalding joined EMBL-EBI in 2011 as part of the Database of Genomic Variants archive (DGVa) team, working with data on structural variation in all species. He became project lead for the European Genome-phenome Archive (EGA) in 2014. The EGA provides the necessary security required to control access, and maintain patient confidentiality, while providing access to those researchers and clinicians authorised to view the data.

Kenneth Haug joined EMBL-EBI in 2009 and is currently a Project Leader in Metabolomics team. Kenneth has previously worked for the DSEG and ChEBI groups at EMBL-EBI. He is a member of the Phenomenal development team.