US-UK PETs Pilot Scheme

**First Draft Use Case 20/3/24 – version 0.04**

Meeting notes agreed items:

Technical deep dive table top session arranged for Thursday 18th

List of initial data providers in scope:

* NDRS (England) data provider
* SEER programme (US) data provider

DARS:

* NHS England (English data only) – Data Access Request Service - Speak to Lucy?
* NHS England Section 251 approval – requires patient have to opt-out (opt-out rates very low).
* SEER 20+ cancer registries across the US covering 48% of the US population – for paediatric data will require further approvals – with additional approvals can take it up to 70%. Single mechanism out to every registry and each individual registry holds their own approval.

Data & Ontologies:

* Data dictionary for all NHS England data.
* Differences between data sets will require normalisation and likely a preferred or agreed standard to which all data should compare and possible transformation rules to get from local to target standard
* Various ML techniques do not require fully formatted equivalent data between data providers for analysis and this should be a stretch function of the next technical table top session
* TP53 gene provides a further use case for genomic federated analysis
* Suggested rule of thumb for 10 years of data on NHS England data sets (2013 UK registry structural data improvement)
* Incomplete data points on different datasets may result in incomplete queries
* Diagnoses to follow ENCR Rules though different practices in different countries
* Discussion on staging definitions – assumption to use Toronto staging enumeration
* Some UK genomic data is stored in excel format
* US requirement for anonymisation to not allow reidentifications of individual labs from data so any provider markers need to be removed
* Most NHS England data will be in XML files rather than FHIR
* Data, such a deprivation index, would require bespoke analysis for comparison
* No equivalent definition of rurality in England
* US has 3k+ counties and rurality defined at county levels
* Will require multiple technical sessions for standardisation / agreement on enums / terms
* Requirement to capture outcome status (migrated, survival, recurrence) with standardisation of enumerated values

Synthetic Data:

* Will use synthetic data for all pre-production environments and testing
* Simulacrum generates cancer data <https://simulacrum.healthdatainsight.org.uk/>
* Synthea is an open source tool https://synthea.mitre.org/downloads
* MOSAIC suggested by Betsy Hsu <https://www.air.org/mosaic/tools>
* MOSAIC already runs on top of SEER data (to validate)
* NIST have synthetic data benchmarking tools (to validate)

Future Data Types:

* Initial scope to be limited to the workable with allowance for extensibility
* Genomic sequence data (different file types), imaging data, biopsy and histopathology imagery data will be included if beneficial as multi-modal data and can be used in federated learning scenarios

Questions:

* Can we have some specific statistical methods that both parties would use?
* How can those statistical methods be applied to distributed data sets?
* A number of open questions relating to data formatting, ontologies & standardisation which would require workshops post agreement of data providers.

Updated Use Case following meeting Friday 5th April 2024:

For (1) childhood cancer overall, (2) childhood cancer groupings classified according to the International Classification of Childhood Cancer version 3 (ICCC3) and (3) specific, molecularly-defined central nervous system tumour entities including the molecular subtypes of medulloblastoma with ICD-10 Code (C71.\_) and ICD-O morphology (9470/3, 9471/3, 9473-9477/3)

*[PLEASE AMEND THE FOLLOWING]*

Suggested Research Use cases:

* As a medical researcher I want to,
  + statistically study the disease by making DAR requests against individual territories to enable privacy preserving statistical studies across territories
  + compare data of all childhood cancers combined and individual ICCC3 groupings by sex, ethnicity, deprivation and rurality
  + run federated and aggregated data analytics on distributed territory data sets (e.g Anova, R2)
  + form longitudinal studies of disease to evaluate the relationship between risk factors and the development of disease, and the outcomes of treatments over different lengths of time. I want access to SES data and disease information.
  + examine cohorts of data and associated treatments
  + access the following data types:

Aggregated data:

* Number of diagnosed cases overall and by ICCC3 grouping in territory
* Number/proportion of medulloblastomas that are genetically versus histologically defined
* Definition of size of territory
* Time period-based recovery rates
* Cohorts of patients, treatments & results
* Survival and variations in survival by per-patient metrics below
* Dependent upon research question would be important to compare on countries, survival rates, differential treatment

Per patient data:

* Patient age & sex
* Patient socioeconomic status information
* Patient familial history of the disease\*
* Cancer stage using appropriate cancer-specific staging systems
* Patient residence information (UK postcodes, deprivation quintiles, distance from treatment centres)
* Patient conditions experienced (secondary malignancies)
* CSF cytology result
* Ethnicity & race (allowing for US vs UK definitions & heritage)
* Table of ICC3, Country, Survival Rate, Treatment, Stage (other outcomes), recurrence
* Recurrence
* Date / my / of last treatment (follow up)
* Cause of death
* Subsequent primary cancers at individual level
* Interval times between events (treatment, imaging) and Time To Event
* Cancer stage status (see notes re Toronto)

Version History

V0.02 – Updates from Brian Rous & Martin McCabe

V0.03 – Updates from Anna Squicciarini & Heide Hanson

V0.04 - Updates from meeting 5th April including agreed items + updated use case + formatting