Building cancer cohorts

Emily Perry

Research Engagement Manager

14th April 2026

Data security

- This training session will include data from the GEL Research Environment
- As part of your IG training you have agreed to not distribute these data in any way
- If you are joining virtually, you are not allowed to:
    - Invite colleagues to watch this training with you
    - Take any screenshots or videos of the training
    - Share your webinar link (we will remove anyone who is here twice)

Questions

All your microphones are muted

Use the Zoom Q&A to ask questions

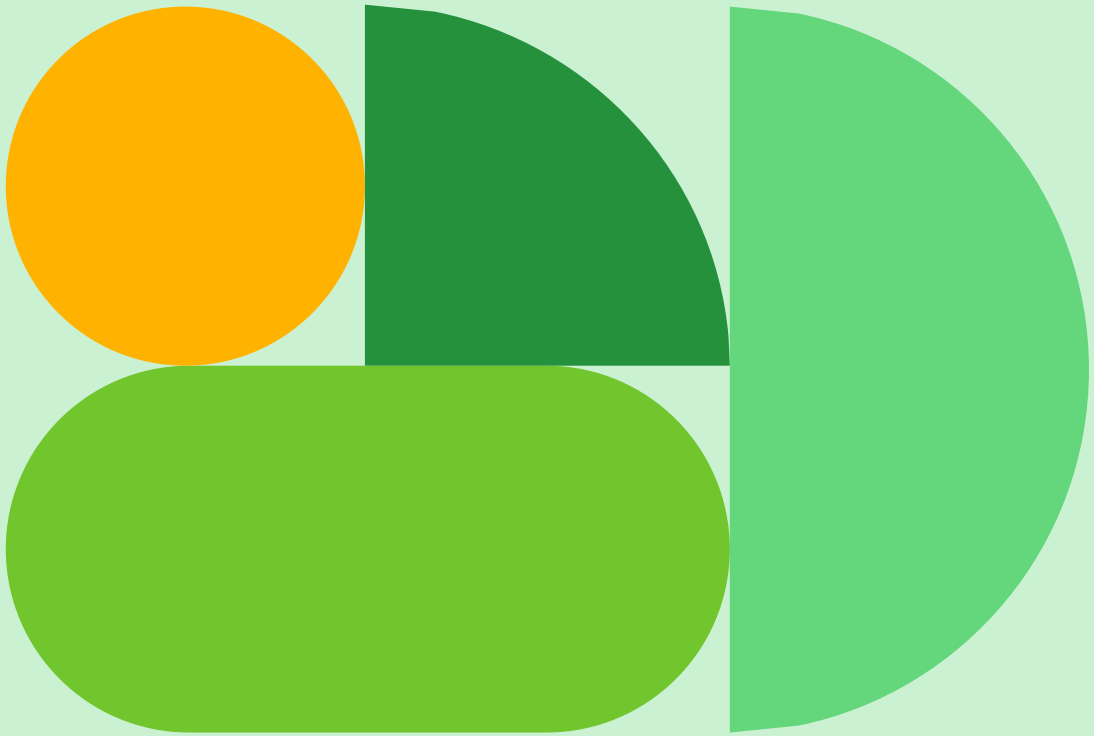
Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Helpers

Matthieu Vizuete-Forster
Learning designer

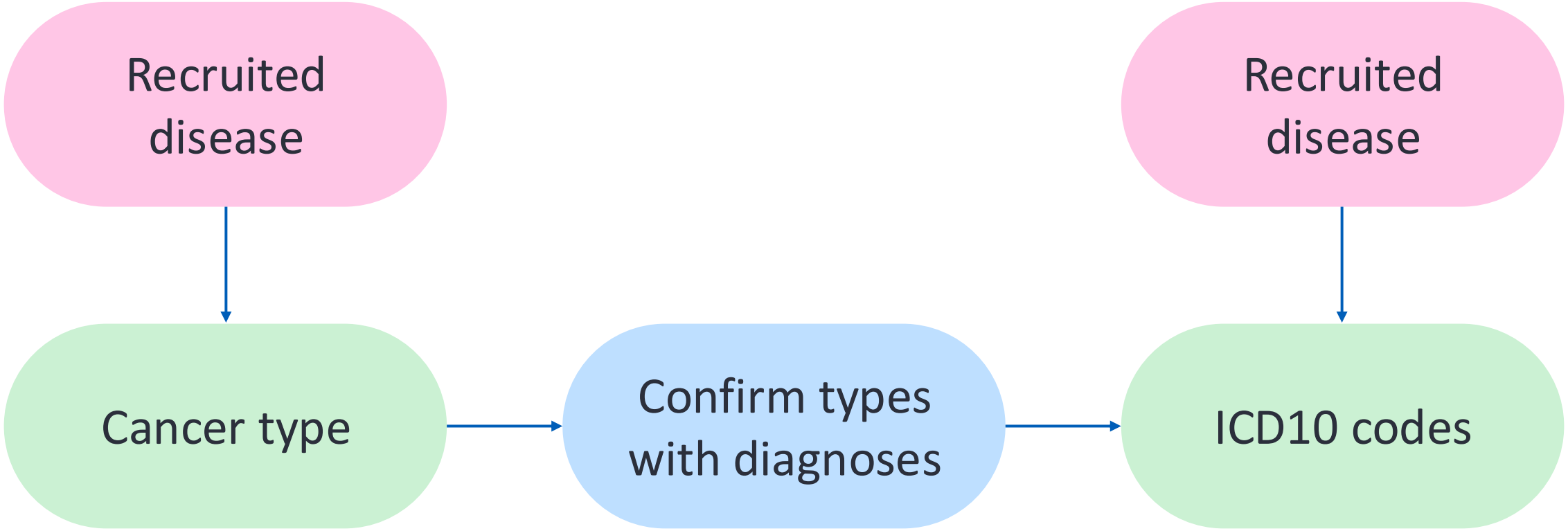
Agenda

1. Introduction and admin
2. Parameters and considerations for building a cohort
3. Point-and-click cohort building with Participant Explorer
4. Tables for cohort building in cancer
5. Programmatic cohort building in Python and R
6. Getting genomic filepaths for your cohort
7. Using your cohort with aggregate VCFs
8. Help and questions

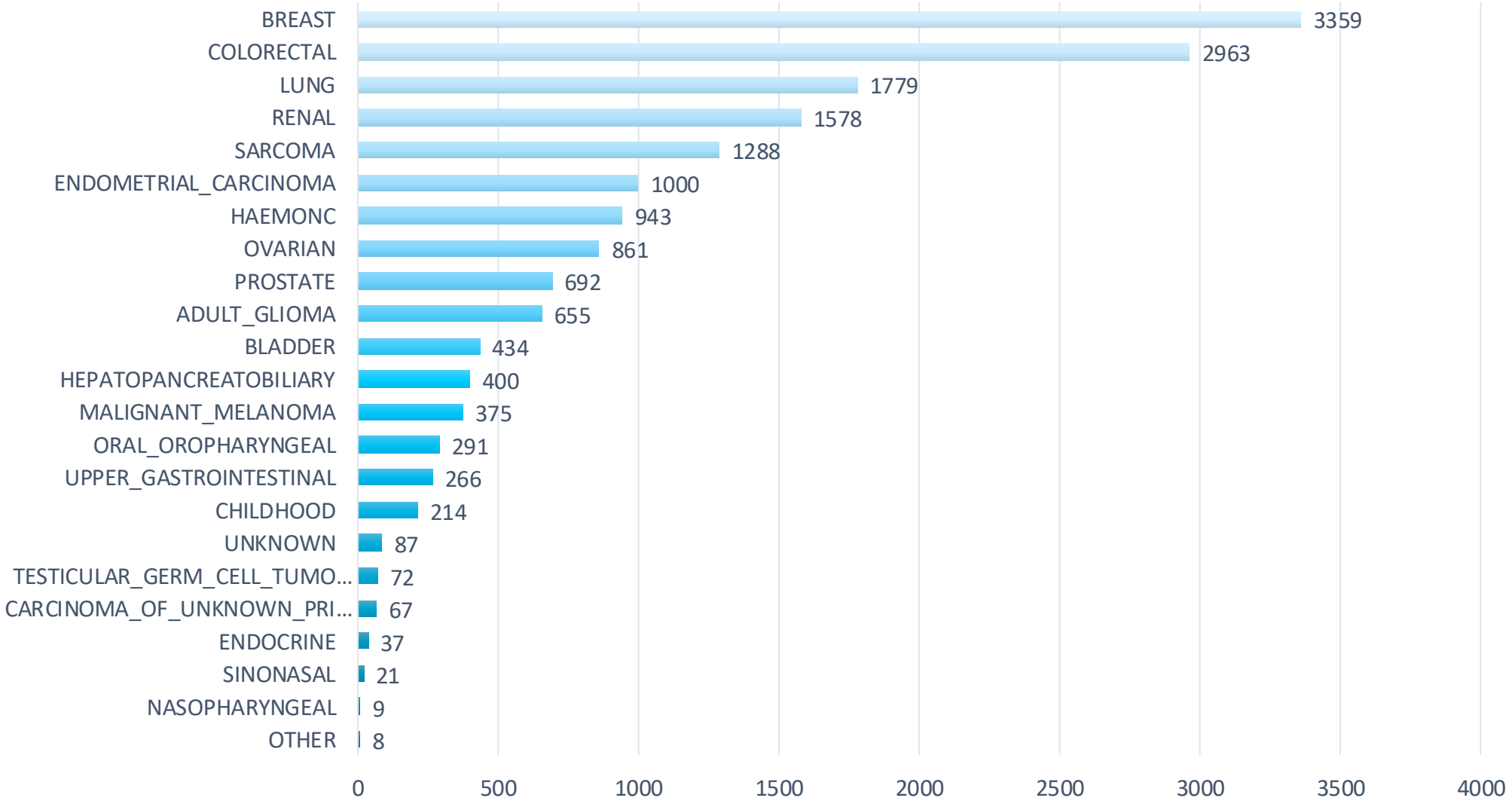


2. Parameters and considerations for building a cohort

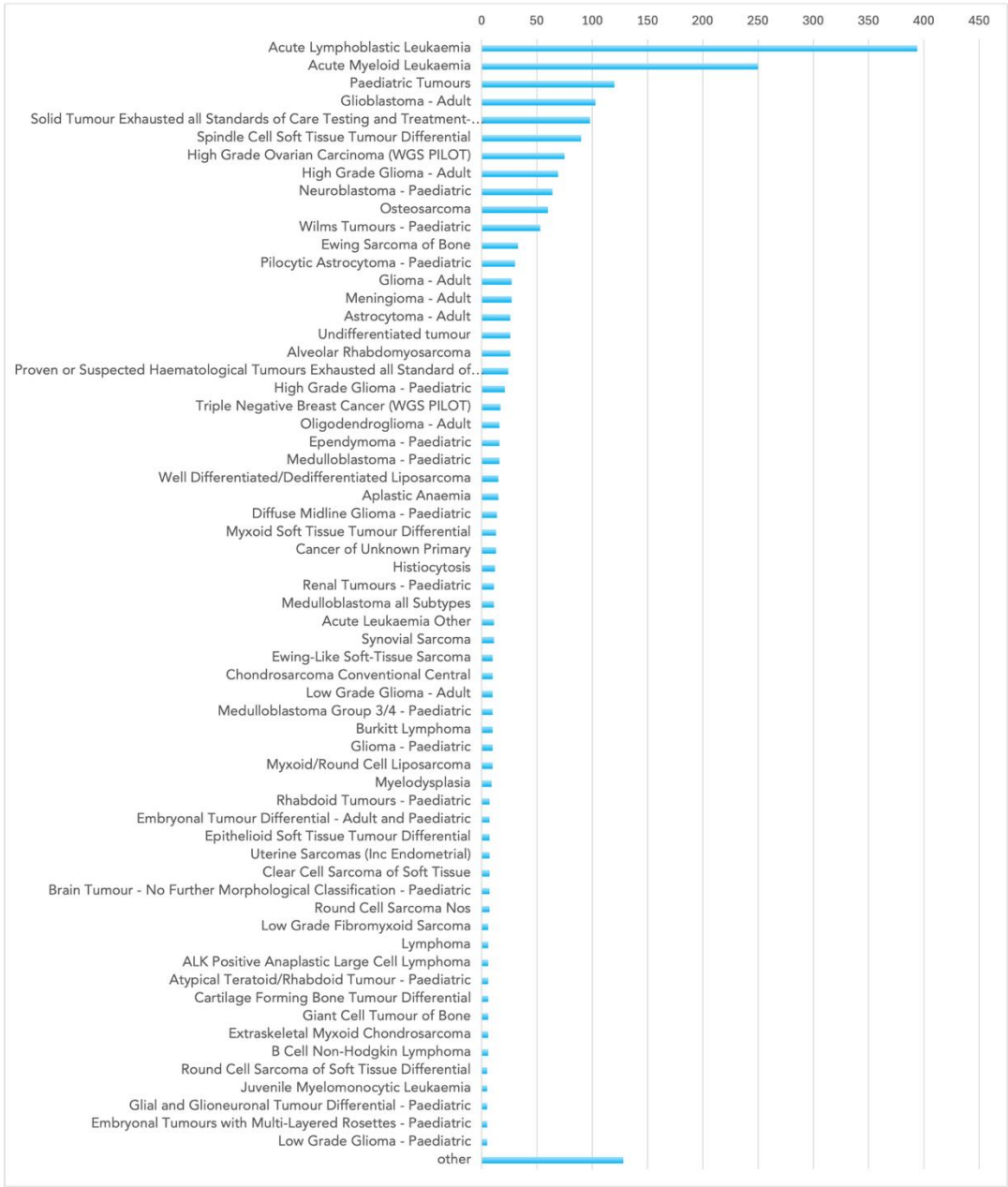
Cancer type


Recruited disease
Recruited disease
Cancer type
Confirm types with diagnoses
ICD10 codes

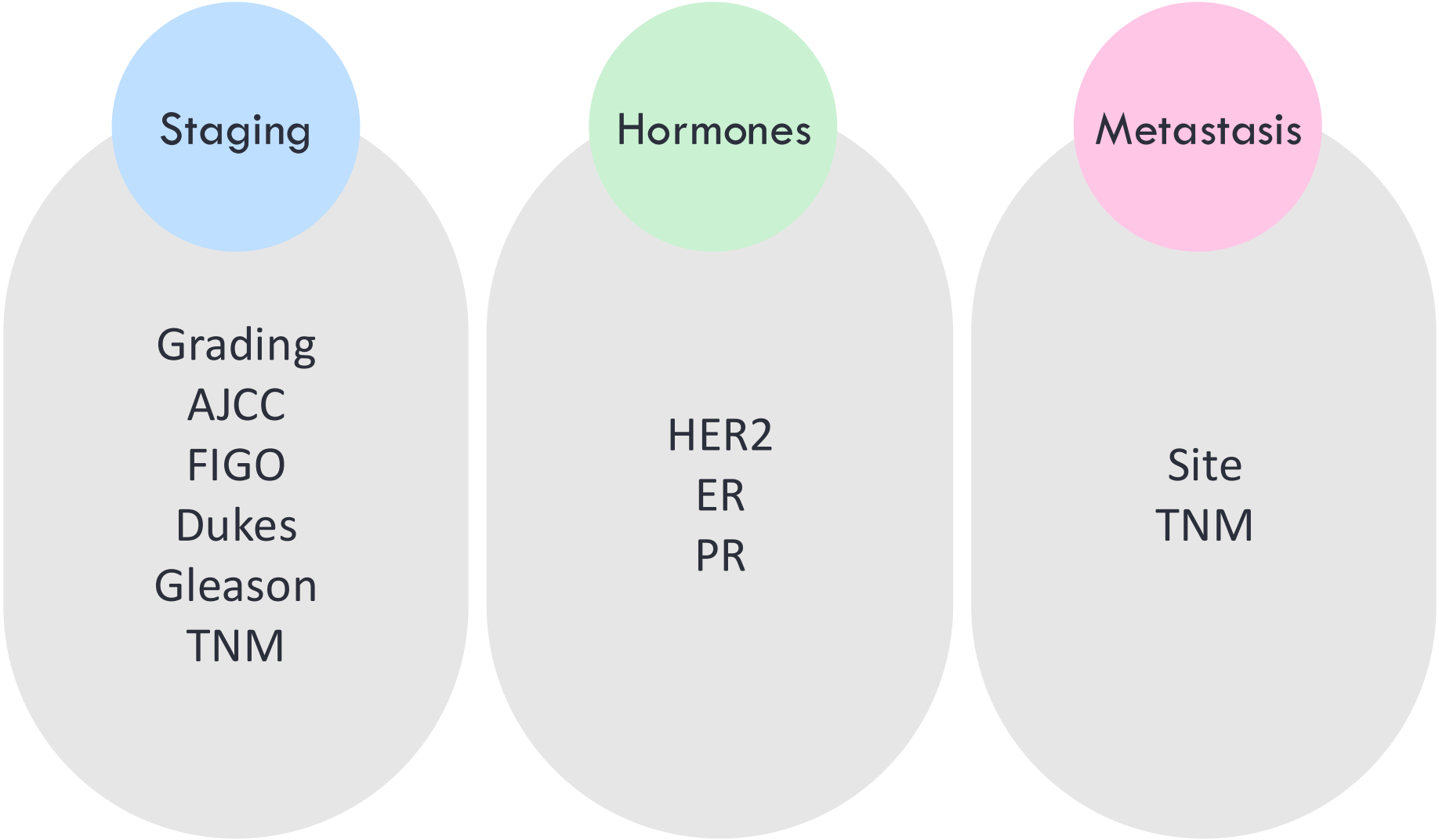
100,000 Genomes cancer



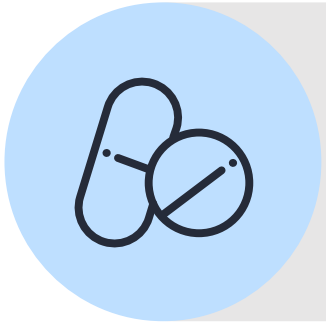
NHS GMS cancer



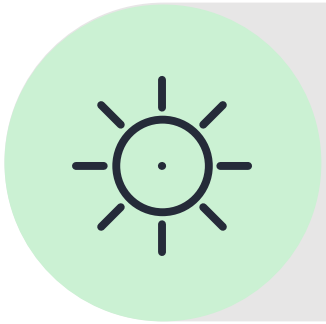
Cancer characteristics



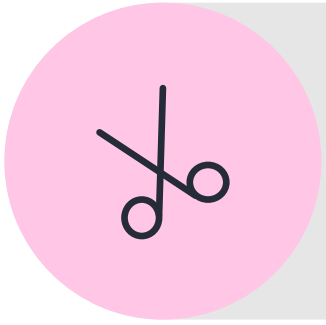
Cancer treatment



Chemotherapy
Immunotherapy

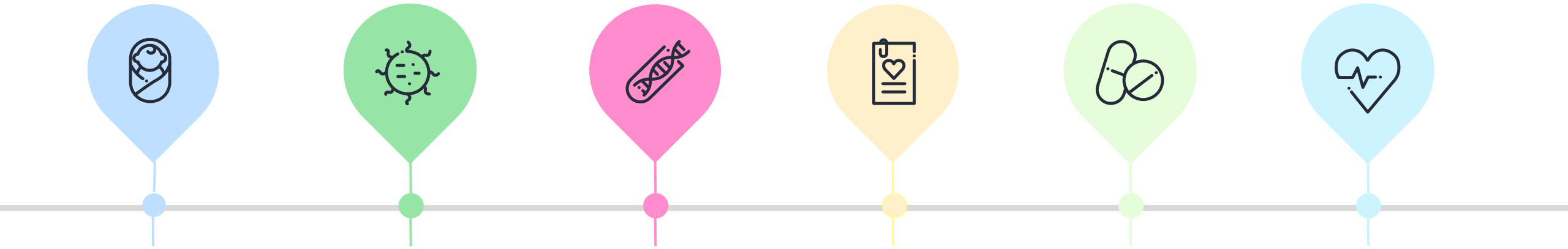


Radiotherapy



Surgery

Derive age from dates



Year of birth

Diagnosis

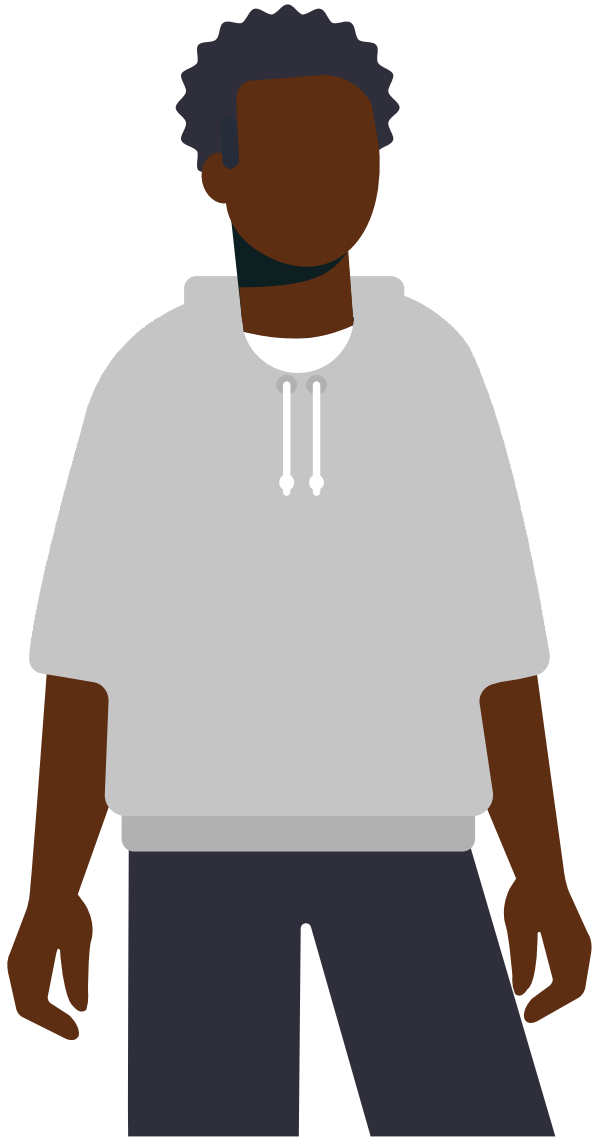
Sampling

Phenotyping

Treatment

Death

Demographics



Index of multiple deprivation

Ethnicity

Smoking

Testicular cryptorchidism

HRT

Endometriosis

Menarche/menopause

Sunbed use

Pregnancy/breastfeeding

IUD

Dialysis

IVF

Hepatitis/cirrhosis

Height/weight

Alcohol

Tamoxifen

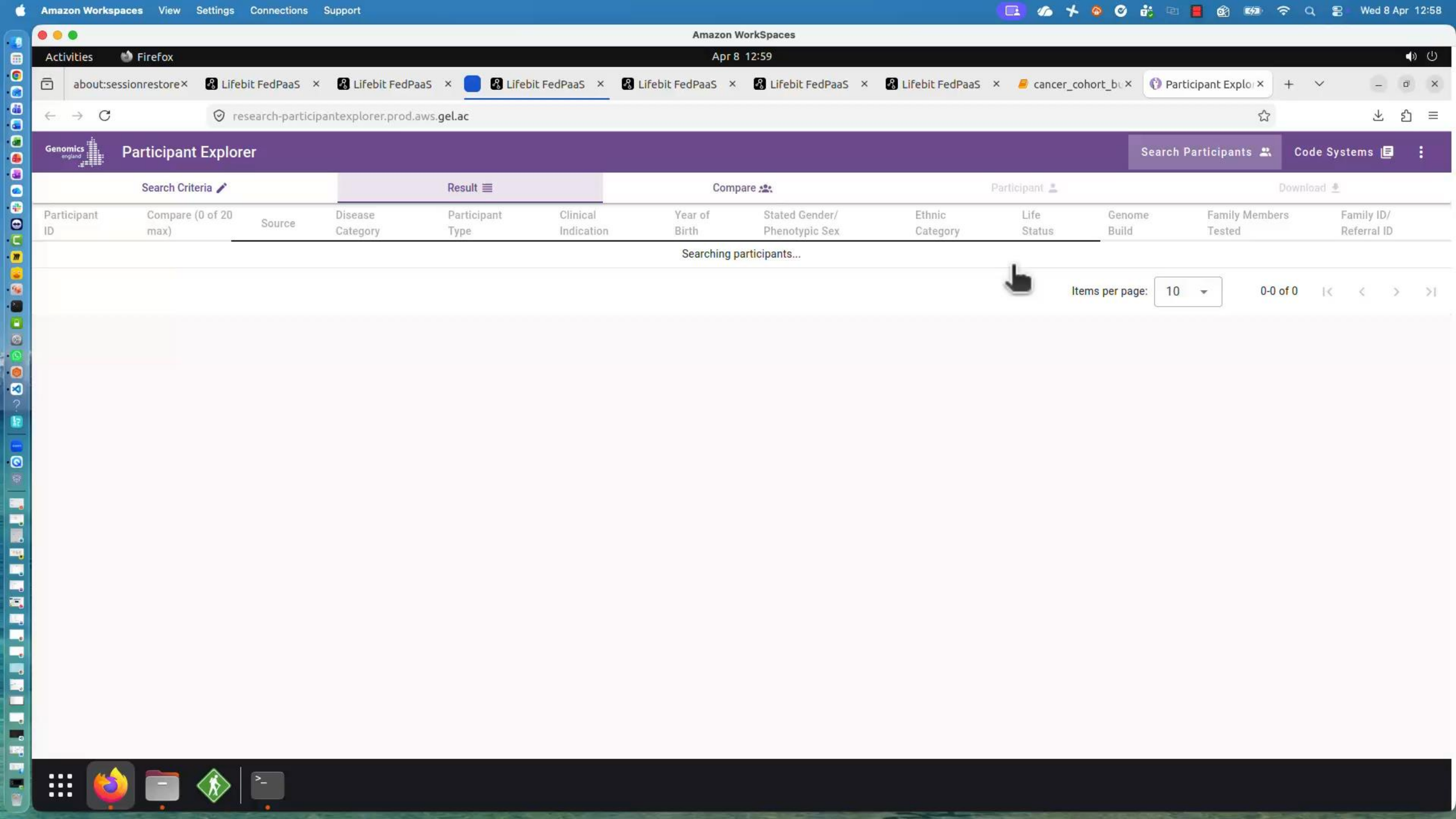
3. Point-and-click cohort building with Participant Explorer

Participant Explorer

- Search for participants by:
  - IDs
  - Clinical concepts
    - diagnoses
    - treatments
    - ontology-aware
  - Personal details
- View/compare medical histories

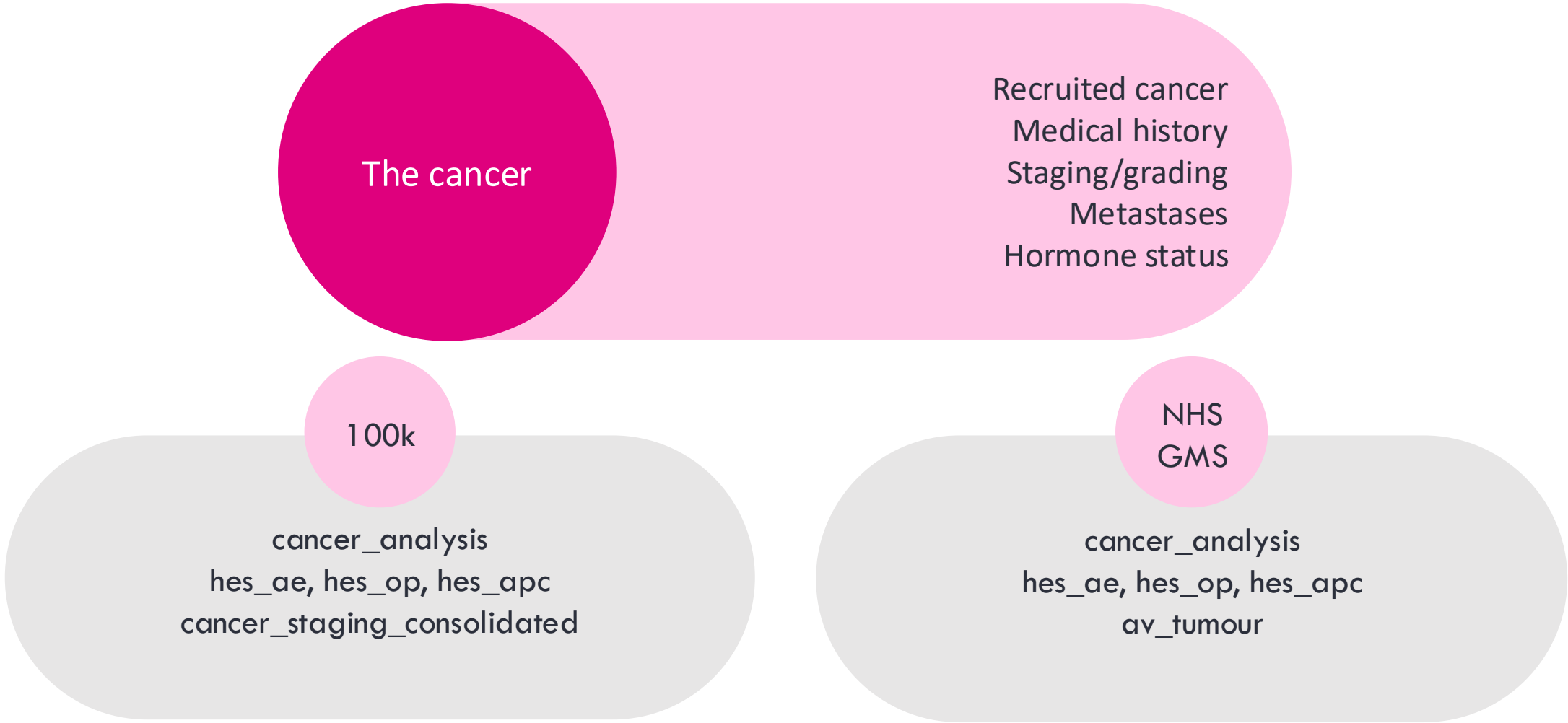
https://re-docs.genomicsengland.co.uk/pxa/

Participant Explorer demo

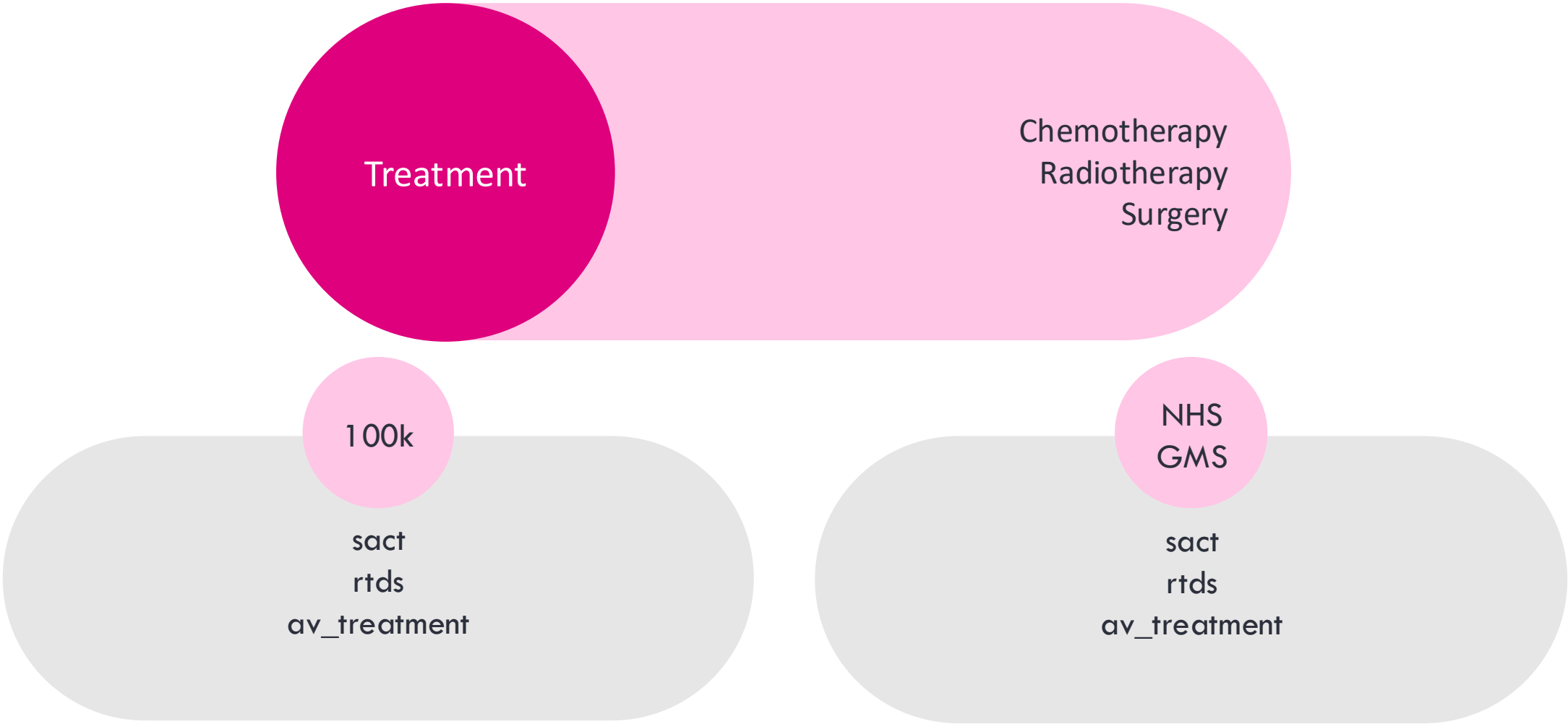

Amazon Workspaces View Settings Connections Support
Wed 8 Apr 12:58
Amazon WorkSpaces
Activities
Firefox
Apr 8 12:59
about:sessionrestore
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
cancer_cohort_b
Participant Explo
research-participantexplorer.prod.aws.gel.ac
Genomics england
Participant Explorer
Search Participants
Code Systems
Search Criteria
Result
Compare
Participant
Download
Participant ID
Compare (0 of 20 max)
Source
Disease Category
Participant Type
Clinical Indication
Year of Birth
Stated Gender/ Phenotypic Sex
Ethnic Category
Life Status
Genome Build
Family Members Tested
Family ID/ Referral ID
Searching participants...
Items per page:
10
0-0 of 0

4. Tables for cohort building in cancer

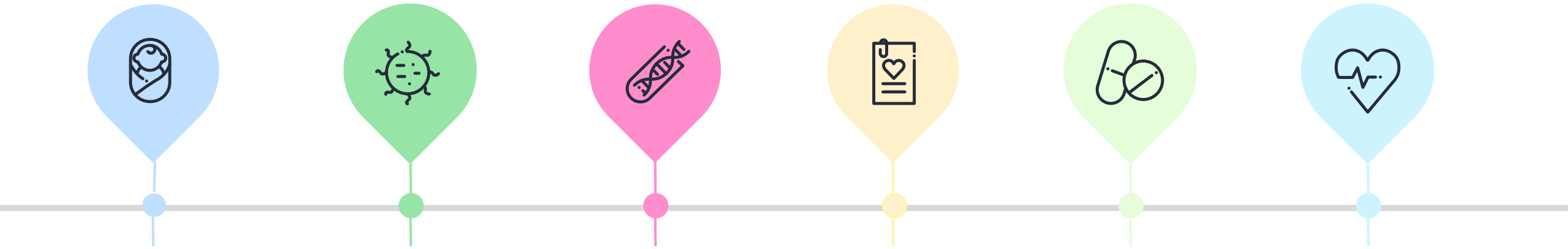
Cancer cohort parameters



Cancer cohort parameters

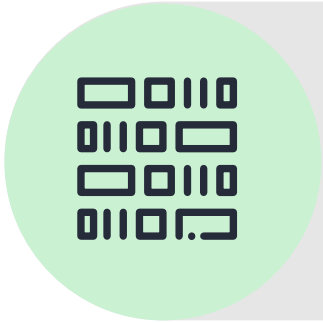


Derive age from dates

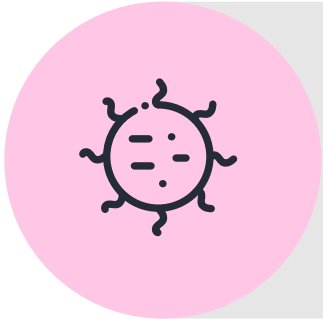


Keys

participant_id

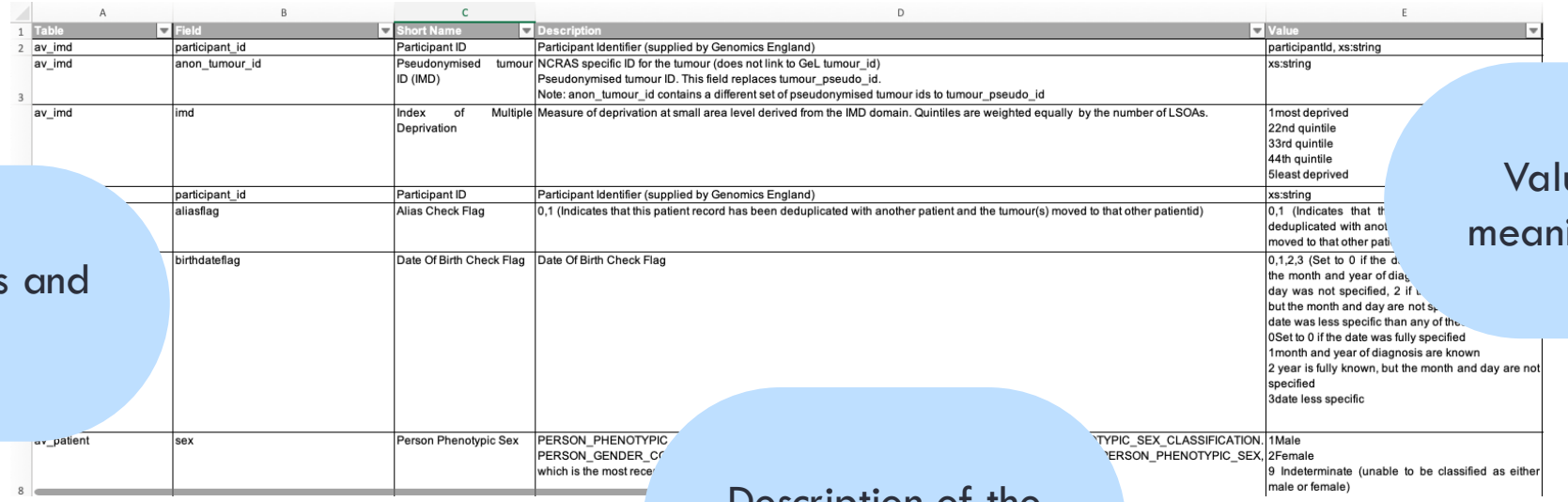


platekey



anon_tumour_id

Data dictionary



	A	B	C	D	E
1	Table	Field	Short Name	Description	Value
2	av_imd	participant_id	Participant ID	Participant Identifier (supplied by Genomics England)	participantId, xs:string
3	av_imd	anon_tumour_id	Pseudonymised tumour ID (IMD)	NCRAS specific ID for the tumour (does not link to GeL tumour_id) Pseudonymised tumour ID. This field replaces tumour_pseudo_id. Note: anon_tumour_id contains a different set of pseudonymised tumour ids to tumour_pseudo_id	xs:string
	av_imd	imd	Index of Multiple Deprivation	Measure of deprivation at small area level derived from the IMD domain. Quintiles are weighted equally by the number of LSOAs.	1most deprived 22nd quintile 33rd quintile 44th quintile 5least deprived
		participant_id	Participant ID	Participant Identifier (supplied by Genomics England)	xs:string
		aliasflag	Alias Check Flag	0,1 (Indicates that this patient record has been deduplicated with another patient and the tumour(s) moved to that other patientid)	0,1 (Indicates that th deduplicated with anot moved to that other pati
		birthdateflag	Date Of Birth Check Flag	Date Of Birth Check Flag	0,1,2,3 (Set to 0 if the d the month and year of dia day was not specified, 2 if but the month and day are not s date was less specific than any of th 0Set to 0 if the date was fully specified 1month and year of diagnosis are known 2 year is fully known, but the month and day are not specified 3date less specific
8	av_patient	sex	Person Phenotypic Sex	PERSON_PHENOTYPIC PERSON_GENDER_C which is the most rece	TYPIC_SEX_CLASSIFICATION. ERSON_PHENOTYPIC_SEX, 1Male 2Female 9 Indeterminate (unable to be classified as either male or female)

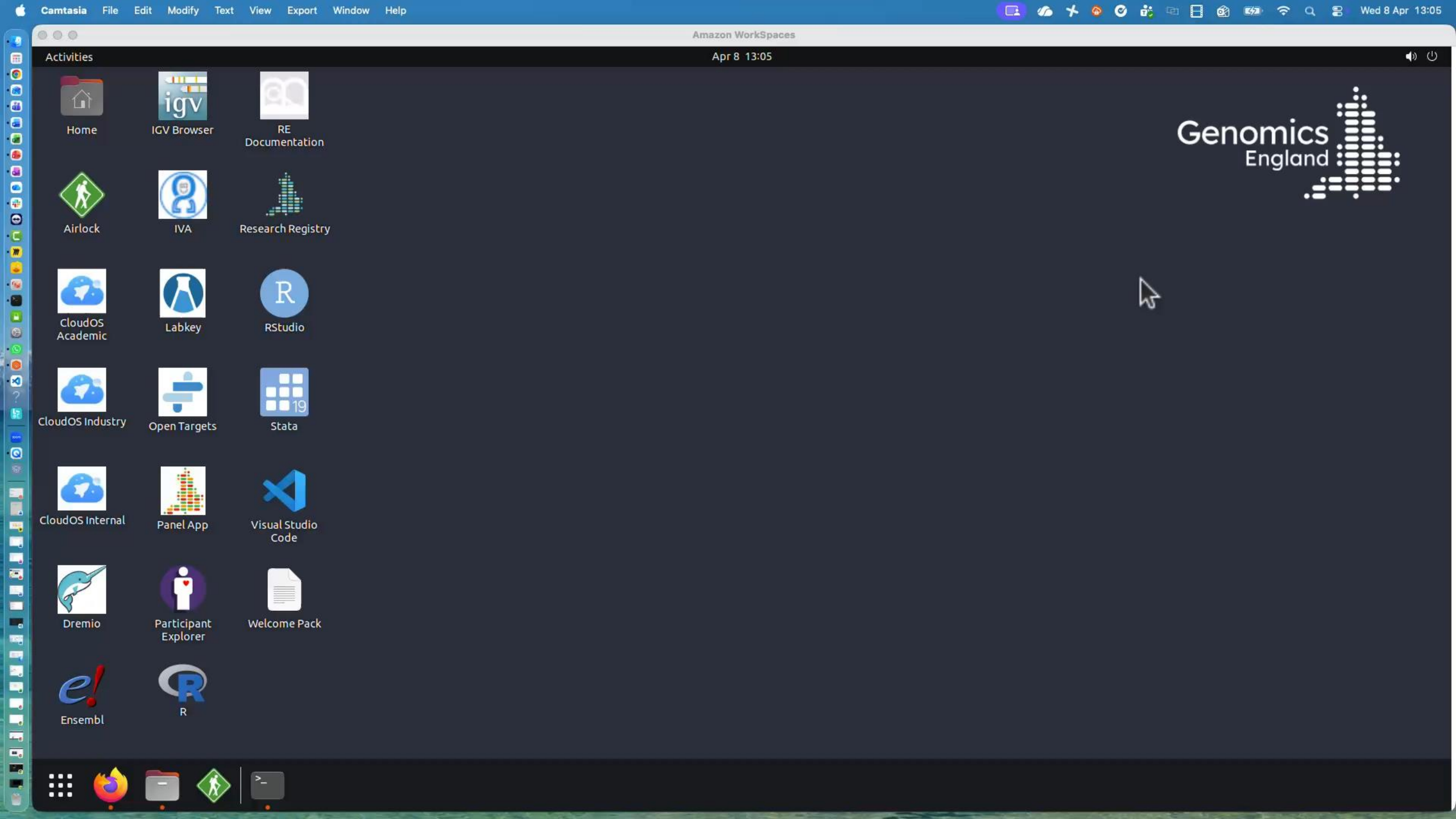
Lists of tables and columns

Value type or meaning of codes

Description of the data

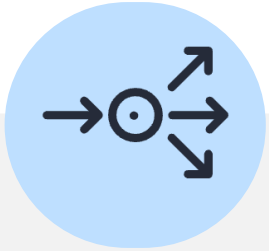
https://re-docs.genomicsengland.co.uk/100k_data_dictionary https://re-docs.genomicsengland.co.uk/GMS_data_dictionary

Tables demo

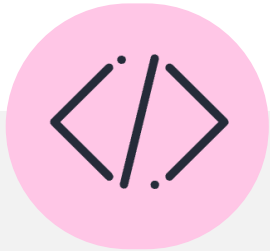

Camtasia File Edit Modify Text View Export Window Help
Wed 8 Apr 13:05
Amazon WorkSpaces
Activities
Apr 8 13:05
Home
IGV Browser
RE Documentation
Airlock
IVA
Research Registry
CloudOS Academic
Labkey
RStudio
CloudOS Industry
Open Targets
Stata
CloudOS Internal
Panel App
Visual Studio Code
Dremio
Participant Explorer
Welcome Pack
Ensembl
R
Genomics England

5. Programmatic cohort building in Python and R

LabKey API



Combine queries between tables



Work in a variety of programming languages (support for Python and R) using SQL queries

Replicate queries between releases and analyses

Work locally and on the HPC

https://re-docs.genomicsengland.co.uk/labkey_api/

LabKey .netrc

- You can access the same data via the LabKey API as you can through other means
- You will need to configure access to the LabKey API with your username and password
    - In your home directory
    - On the HPC
- You do this by editing a file called .netrc

https://re-docs.genomicsengland.co.uk/labkey_api/

Programming tools in the RE

https://re-docs.genomicsengland.co.uk/hpc_jupyter/ https://re-docs.genomicsengland.co.uk/enable_rstudio/

Demo notebooks


```
/gel_data_resources/example_scripts/
workshop_scripts/cancer_cohorts_2026
```

Programming demo

Amazon Workspaces View Settings Connections Support

Wed 8 Apr 13:17

Amazon WorkSpaces

Activities Firefox Apr 8 13:17

cancer_cohort_building_

http://localhost:8283/notebooks/cancer_cohorts_2026/cancer_cohor

jupyter cancer_cohort_building_training Last Checkpoint: 7 days ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

Cancer cohort building in Python

This notebook will walk you through building cancer cohorts using the LabKey API in Python. You are welcome to copy/paste any code from this notebook for your own scripts/notebooks.

Contents:

Import Python modules you need

pro.cloud-os.prod.aws.gel.ac/app/interactive-analysis/running/69b2a

cohort building in I Go to Session Save Pause

Add cost limit Time left: 01h 16m 17s

Monitor Usage

Rstudio Session Status: Running

ID: 69b2a564de817422fd3d6394

Started: 03/12/2026 11:37:08

Instance type: c5.xlarge 4 CPUs / 8 GiB

Execution platform: aws

Project name: Emily_test

Last time stopped: 04/07/2026 13:44:49

R version: 4.5.2

Cost: $2.9905

Last time saved: 04/08/2026 13:02:19

Overall duration: 10h 18m 11s

Last saved on 04/08/2026 13:02:19

Input Data Add data

Linked folders No data

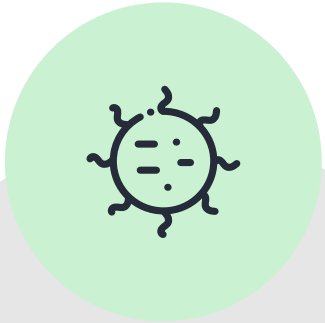
Data Items (5)

6. Getting genomic filepaths for your cohort

Genomic files available

Germline

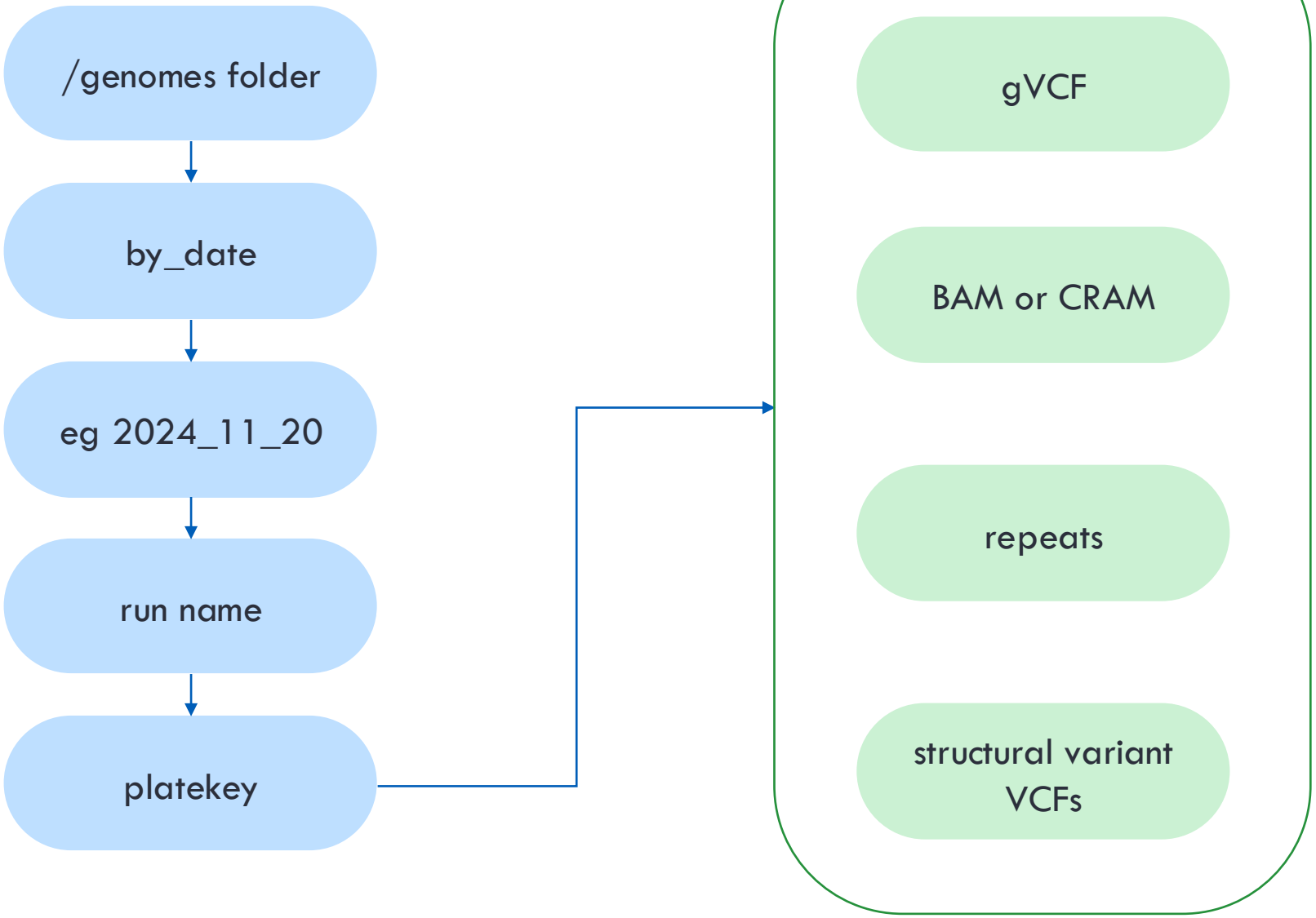
Platekey
BAM
gVCF
SV VCF



Tumour

Platekey
BAM
gVCF
SV VCF

Genomes location



https://re-docs.genomicsengland.co.uk/genomic_data/

File locations

Participant ID	Platekey	Filepath	Filename	File sub-type
12345678	LP12345-DNA_A01	/genomes/by_date/2025-04-08/000111222333/LP12345-DNA_A01/Variations/LP12345-DNA_A01.vcf.gz	LP12345-DNA_A01.vcf.gz	Genomic VCF
12345678	LP12345-DNA_A01	/genomes/by_date/2025-04-08/000111222333/LP12345-DNA_A01/Assembly/LP12345-DNA_A01.cram	LP12345-DNA_A01.cram	CRAM
12345678	LP12345-DNA_A01	/genomes/by_date/2025-04-08/000111222333/LP12345-DNA_A01/Variations/LP12345-DNA_A01.SV.vcf.gz	LP12345-DNA_A01.SV.vcf.gz	Structural VCF

Filepaths demo

Camtasia File Edit Modify Text View Export Window Help

Wed 8 Apr 13:44

Amazon WorkSpaces

Activities Firefox Apr 8 13:44

cancer_analysis: /main-pr × cancer_100K_genomes_r × Participant Explorer ×

research-participantexplorer.prod.aws.gel.ac

Genomics england **Participant Explorer**

Search Participants Code Systems

Search Criteria | Result 865 | Compare | Participant | Download

Participant ID	Compare (0 of 20 max)	Source	Disease Category	Participant Type	Clinical Indication	Year of Birth	Stated Gender/ Phenotypic Sex	Ethnic Category	Life Status	Genome Build	Family Members Tested	Family ID/ Referral ID
211000022	☐	100kGP	Cancer		Breast, Ductal	1965	Female	White: British	Alive	GRCh38 (Dragen 3.2.22) GRCh38 (NSv4)		
211000049	☐	100kGP	Cancer		Breast, Ductal	1957	Female	White: British	Alive	GRCh38 (Dragen 3.2.22) GRCh37 (NSv2) GRCh38 (NSv4)		
211000059	☐	100kGP	Cancer		Breast, Lobular	1955	Female	White: British	Alive	GRCh38 (NSv4) GRCh37 (NSv2) GRCh38 (Dragen 3.2.22)		
211000267	☐	100kGP	Cancer		Breast, Ductal	1935	Female	White: British	Deceased	GRCh38 (Dragen 3.2.22) GRCh38 (NSv4)		
211000405	☐	100kGP	Cancer		Breast, Ductal	1941	Female	White: British	Alive	GRCh38 (NSv4) GRCh38 (Dragen 3.2.22)		
211000464	☐	100kGP	Cancer		Breast, Lobular	1974	Female	Asian or Asian British: Indian	Alive	GRCh38 (NSv4) GRCh38 (Dragen 3.2.22)		
211000492	☐	100kGP	Cancer		Breast, Ductal	1938	Female	White: British	Alive	GRCh38 (Dragen 3.2.22) GRCh38 (NSv4)		
211000532	☐	100kGP	Cancer		Breast, Ductal	1961	Female	White: British	Alive	GRCh38 (Dragen 3.2.22) GRCh38 (NSv4)		
211000768	☐	100kGP	Cancer		Breast, Ductal	1935	Female	White: British	Alive	GRCh38 (Dragen 3.2.22) GRCh38 (NSv4)		

7. Using your cohort with aggregate VCFs

Aggregate VCFs

Combined gVCFs

Quality metrics

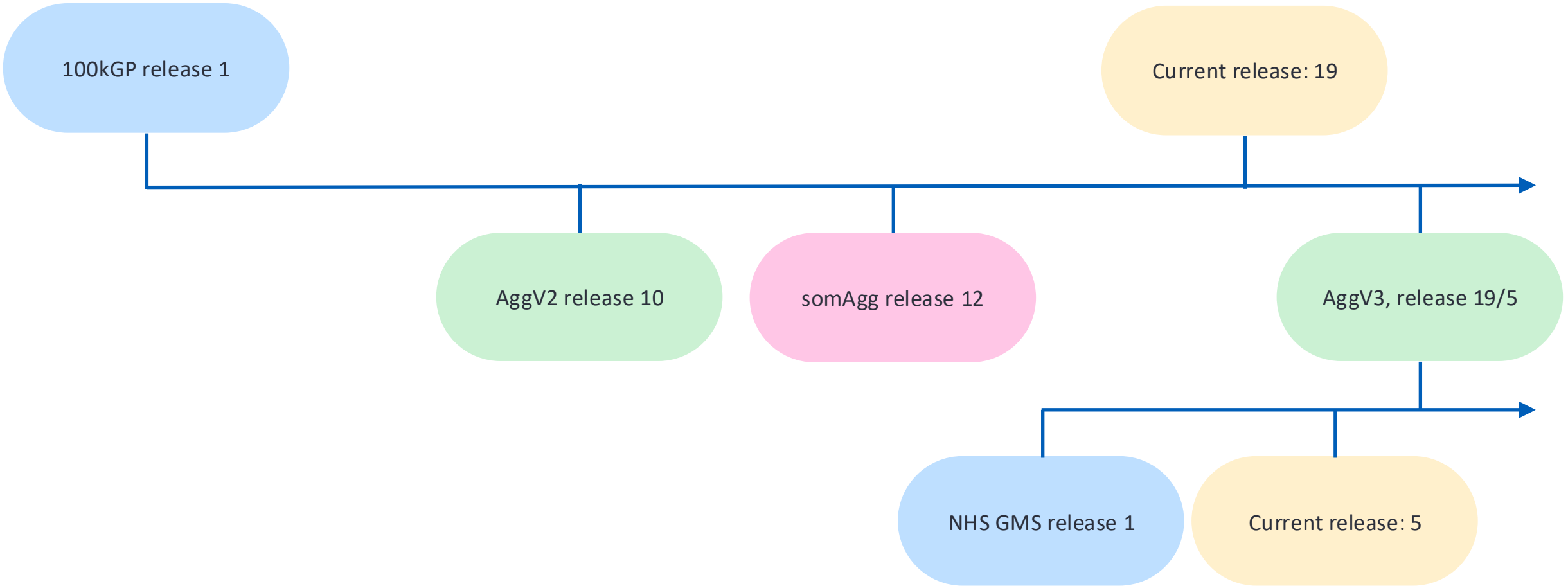
Functional annotation files

Genotype files – platekeys

Chunks/ shards

https://re-docs.genomicsengland.co.uk/aggv3/ https://re-docs.genomicsengland.co.uk/aggv2/ https://re-docs.genomicsengland.co.uk/somAgg/

Aggregate VCFs



https://re-docs.genomicsengland.co.uk/aggv3/ https://re-docs.genomicsengland.co.uk/aggv2/ https://re-docs.genomicsengland.co.uk/somAgg/

Aggregate VCFs

Using bedtools to identify relevant chunks of the aggregate files

Excluding participants who have since withdrawn consent

Using bcftools to extract data from the aggregate VCFs

Codebooks

https://re-docs.genomicsengland.co.uk/aggv2_code_book/ https://re-docs.genomicsengland.co.uk/somAgg_code_book/ https://re-docs.genomicsengland.co.uk/aggv3_code_book/

Aggregate demo

561	pp162723306570	LP5000453-DNA_F01	/genomes/by_date/2021-09-23/BE00043923/LP50004...	2087
562	pp146290195680	LP5100281-DNA_F01	/genomes/by_date/2022-12-01/BE00102497/LP51002...	2089
563	pp109435850368	LP5000773-DNA_G02	/genomes/by_date/2022-06-21/BE00084758/LP50007...	2090
564	pp154791227370	LP5000235-DNA_H01	/genomes/by_date/2021-06-08/BE00029410/LP50002...	2091
565	pp191283296836	LP5100444-DNA_D01	/genomes/by_date/2023-04-12/BE00117387/LP51004...	2095

566 rows × 4 columns

We can write these to a file to use for further analysis.

[70]:
```
path_query.to_csv('python_100k_paths.csv', index=False)
gms_path_query.to_csv('python_gms_paths.csv', index=False)
```

Working with aggregate VCFs

We can query the aggregate VCFs with our cohorts. Let's start by creating a list of platekeys to use for our querie.

[71]:
```
platekeys = ','.join(path_query['germline_sample_platekey'])
platekeys
```

[71]:
```
'LP3000586-DNA_G02,LP3000794-DNA_A01,LP3000314-DNA_A12,LP3000314-DNA_B02,LP3000314-DNA_E08,LP3000314-DNA_E09,LP3000386-DNA_A06,LP3000432-DNA_B03,LP3000386-DNA_A05,LP3000586-DNA_H07,LP3000398-DNA_C07,LP3000398-DNA_G01,LP3000431-DNA_B11,LP3000432-DNA_F05,LP3000432-DNA_F07,LP3000432-DNA_E07,LP3000432-DNA_E08,LP3000620-DNA_G01,LP3000432-DNA_E09,LP3000432-DNA_G01,LP3000398-DNA_G06,LP3000398-DNA_H12,LP3000410-DNA_F06,LP3000410-DNA_A02,LP3000432-DNA_E10,LP3000432-DNA_E01,LP3000432-DNA_E02,LP3000431-DNA_D10,LP3000444-DNA_A01,LP3000680-DNA_G02,LP3000398-DNA_G03,LP3001368-DNA_C05,LP3000432-DNA_G04,LP3001121-DNA_B08,LP3000398-DNA_C09,LP3001585-DNA_C04,LP3000398-DNA_C04,LP3000408-DNA_G10,LP3000537-DNA_B01,LP3000408-DNA_G11,LP3001368-DNA_A08,LP3000617-DNA_D08,LP3000408-DNA_G02,LP3000497-DNA_G04,LP3000410-DNA_G12,LP3000427-DNA_F10,LP3000620-DNA_B09,LP3000620-DNA_D06,LP3000956-DNA_A01,LP3000616-DNA_H01,LP3000410-DNA_F10,LP3000698-DNA_F01,LP3001368-DNA_B05,LP3000497-DNA_E10,LP3000627-DNA_D11,LP3000627-DNA_D08,LP3000497-DNA_E09,LP3000657-DNA_E02,LP3000620-DNA_C04,LP3000617-DNA_C12,LP3000627-DNA_D06,LP3000394-DNA_A01,LP3000657-DNA_E01,LP3000627-DNA_E04,LP3000627-DNA_F09,LP3000620-DNA_F11,LP3000411-DNA_F01,LP3000617-DNA_E01,LP3000604-DNA_B01,LP3000620-DNA_D10,LP3000425-DNA_A01,LP3000617-DNA_E12,LP3000620-DNA_D09,LP3000667-DNA_A02,LP3000425-DNA_D01,LP3000794-DNA_C01,LP3000651-DNA_D01,LP3000651-DNA_F01,LP3000680-DNA_C03,LP3000794-DNA_H01,LP3000651-DNA_D02,LP3000680-DNA_B04,LP3000680-DNA_B04,LP3000680-DNA_B04,LP3000680-DNA_B04,LP3000680-DNA_D04,LP3000680-DNA_B06,LP3000554-DNA_B09,LP3000667-DNA_H06,LP3000537-DNA_E01,LP3000682-DNA_A02,LP3000757-DNA_C0
```

8. Getting help and questions

Getting help

Check our documentation:
https://re-docs.genomicsengland.co.uk/
Click on the documentation icon in the environment

Contact our Service Desk:
https://jiraservicedesk.extge.co.uk/plugins/servlet/desk

Training sessions

3rd Tuesday every month

Introduction to the RE

21/4

19/5

16/6

Materials from past training all online

https://re-docs.genomicsengland.co.uk/upcoming/

Training sessions

12/5	Building rare disease cohorts
9/6	Finding participants by genotype

https://re-docs.genomicsengland.co.uk/upcoming/

Feedback

Thank you

Visit: https://re-docs.genomicsengland.co.uk/