Building cancer cohorts

Emily Perry

Research Engagement Manager

13th May 2025

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

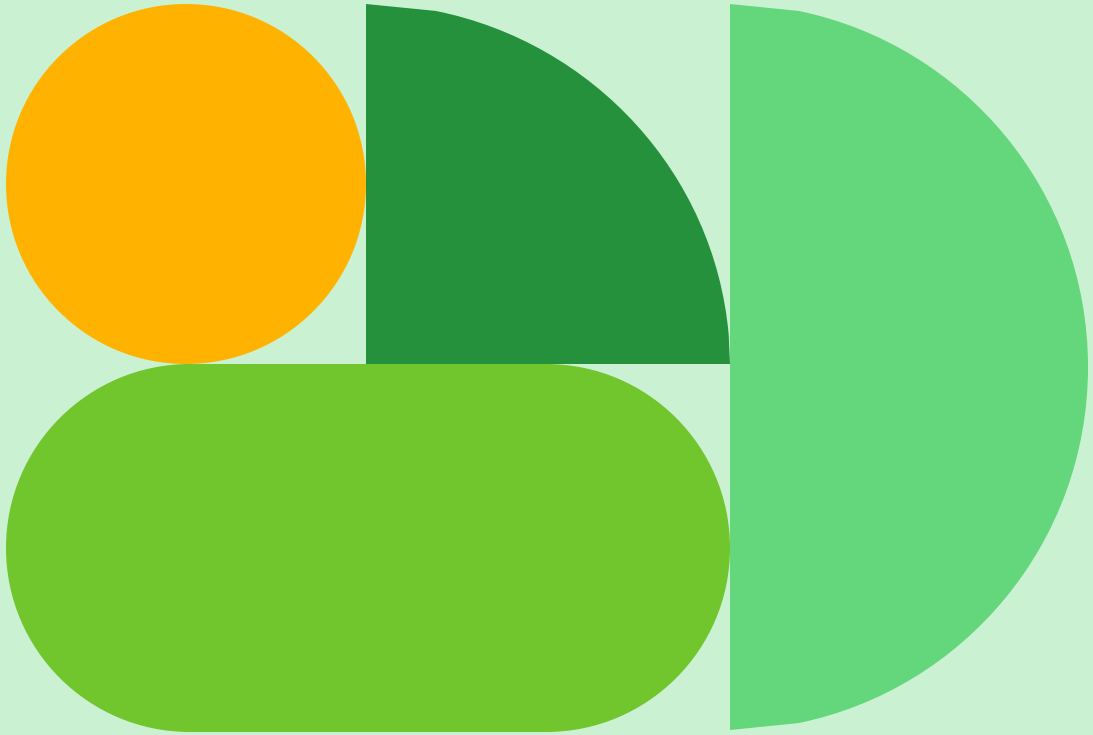
Helpers

Asier Gonzalez-Uriarte
Bioinformatician in Data Release Team (DRT)

Matthieu Vizuete-Forster
Bioinformatician - Research Services

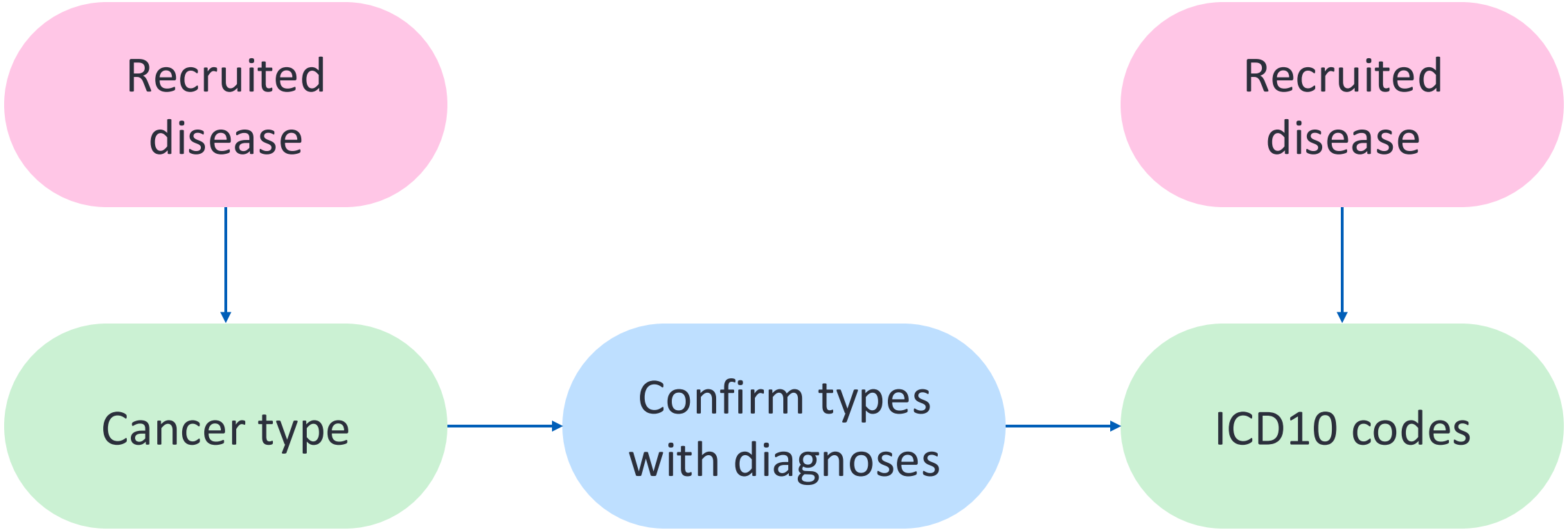
Agenda

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

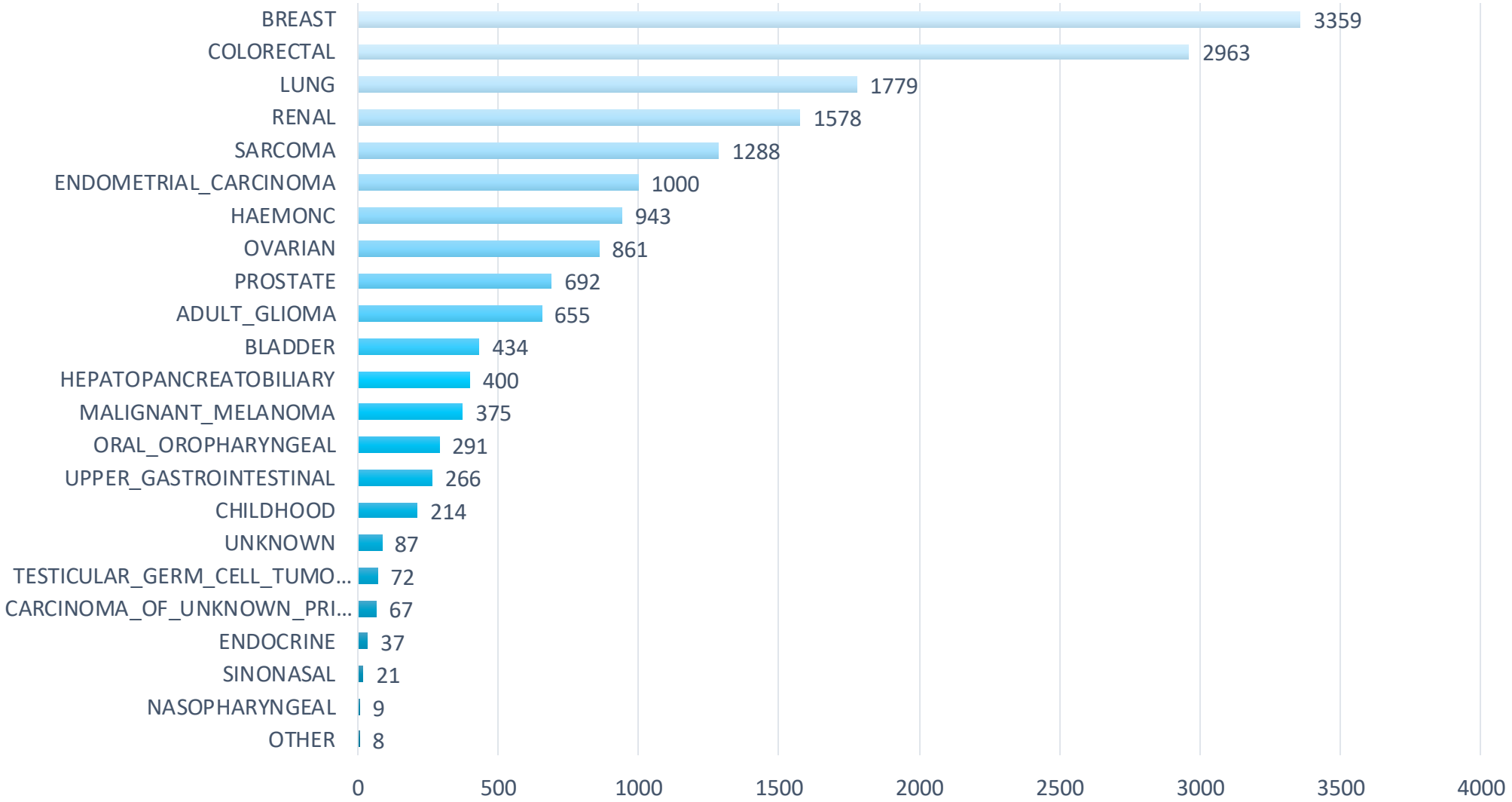


2. Parameters and considerations for building a cohort

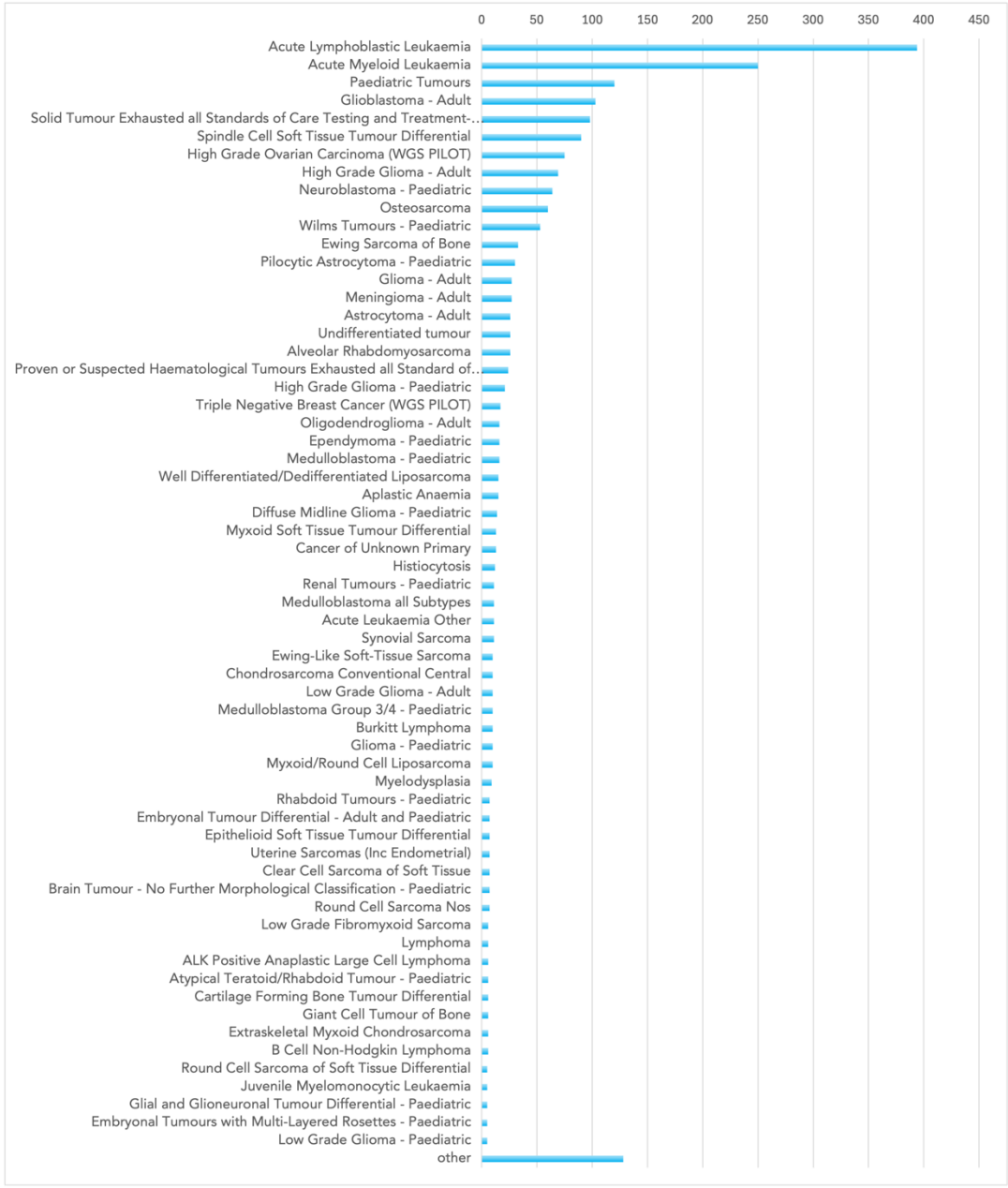
Cancer type


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

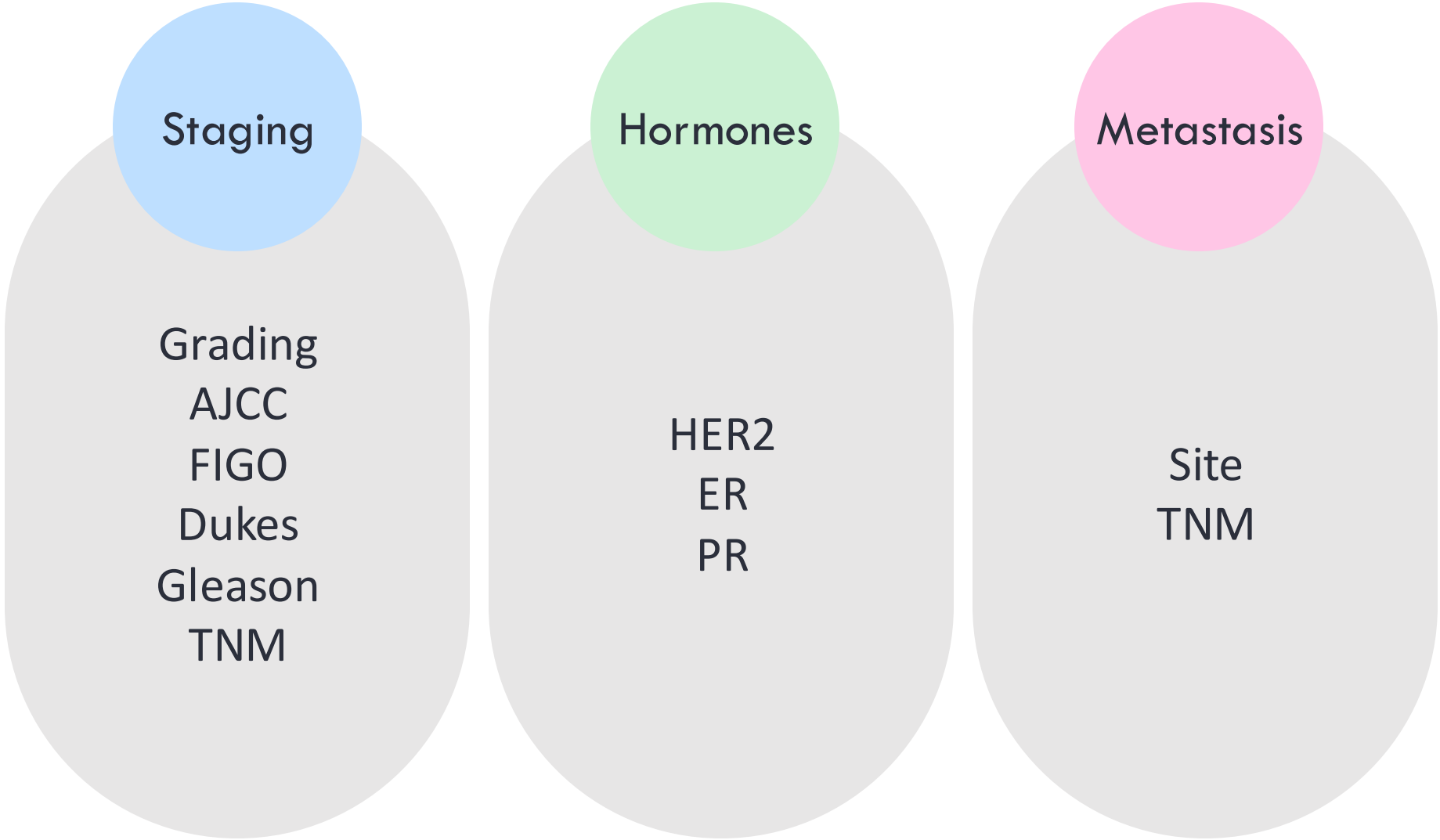
100,000 Genomes cancer



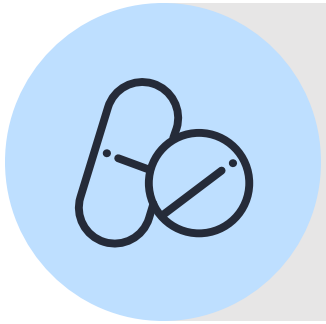
NHS GMS cancer



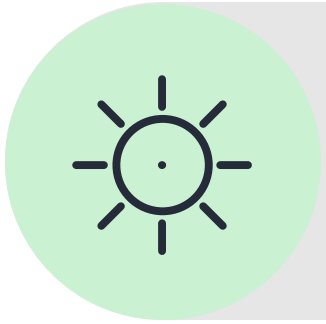
Cancer characteristics



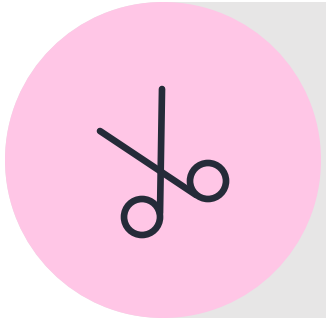
Cancer treatment



Chemotherapy
Immunotherapy

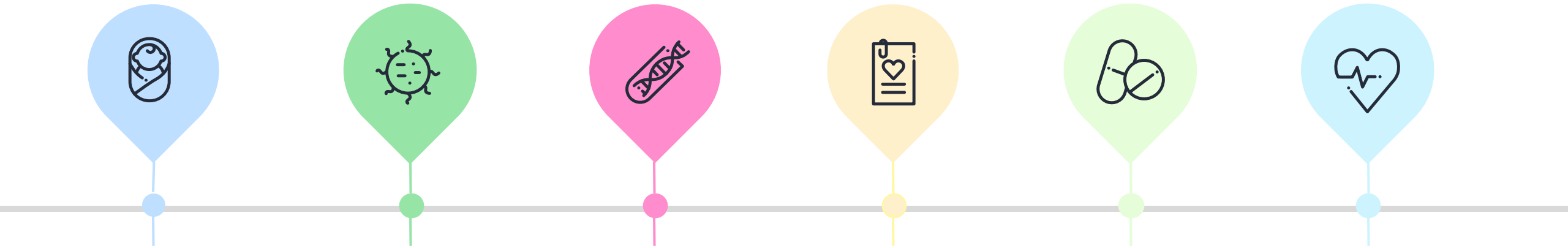


Radiotherapy



Surgery

Derive age from dates



Year of birth

Diagnosis

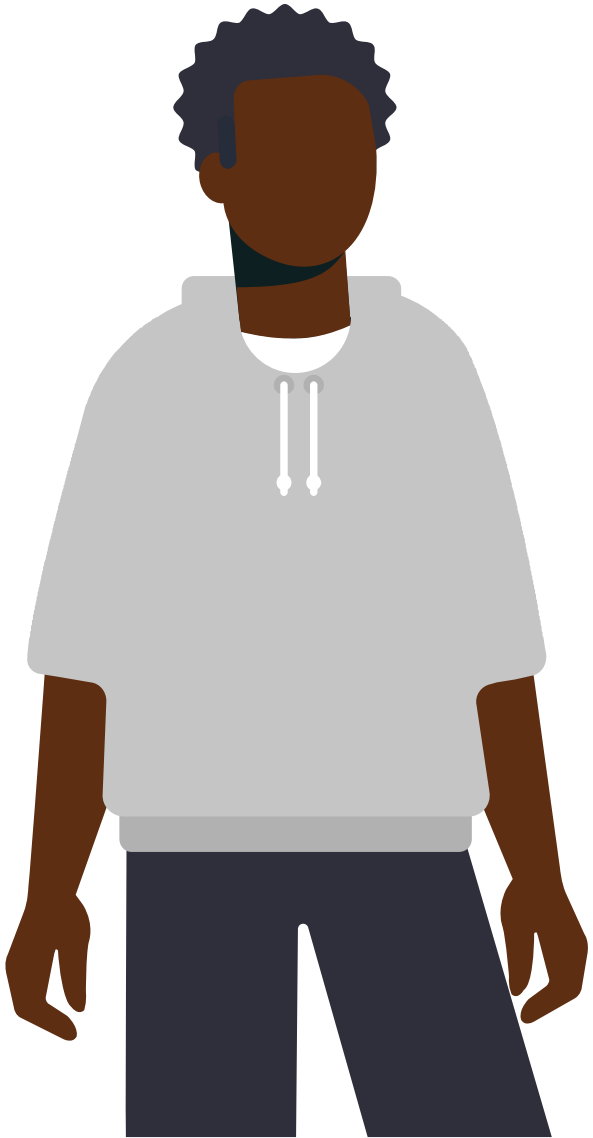
Sampling

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

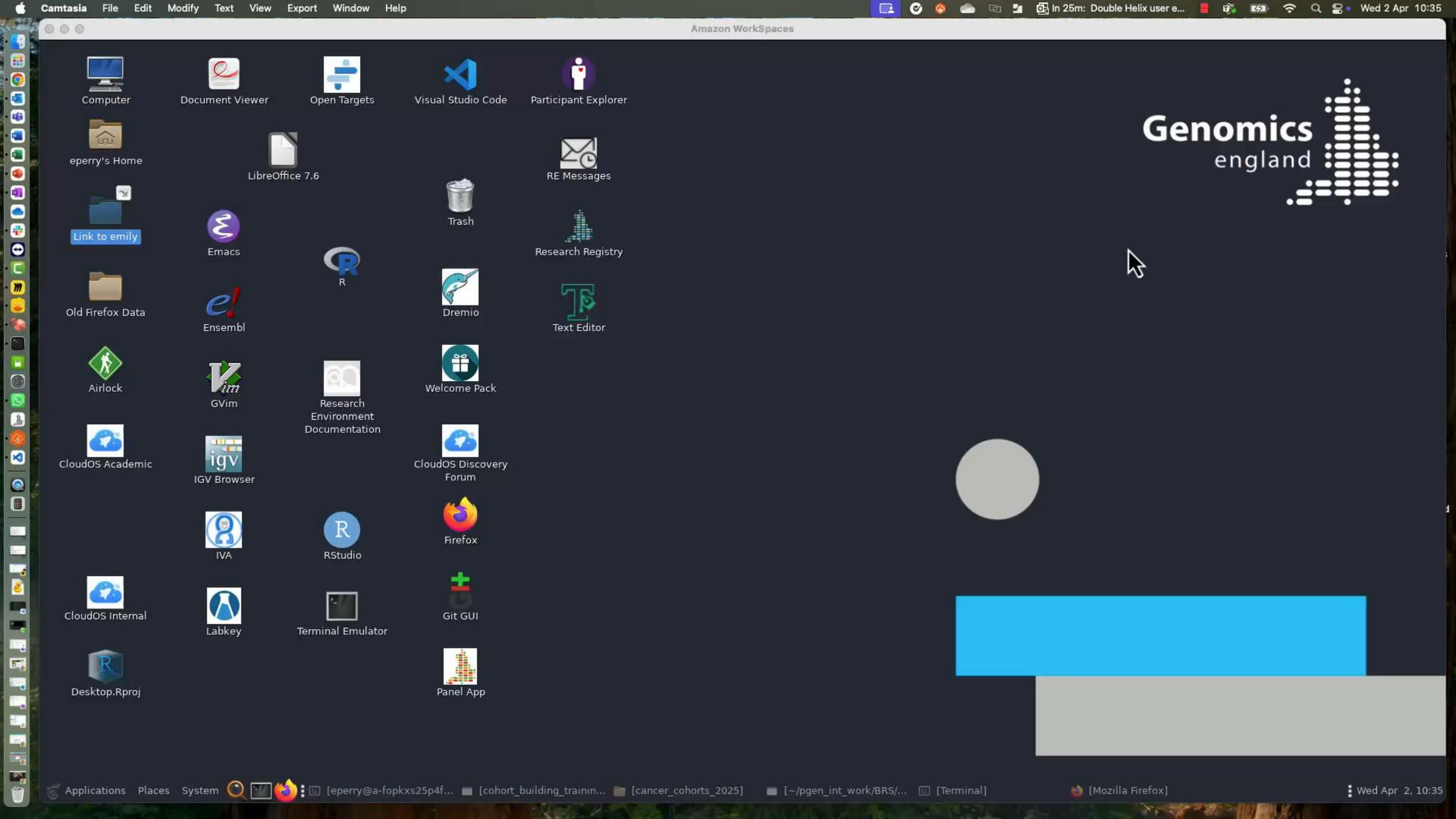
3. Point-and-click cohort building with Participant Explorer

Participant Explorer

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

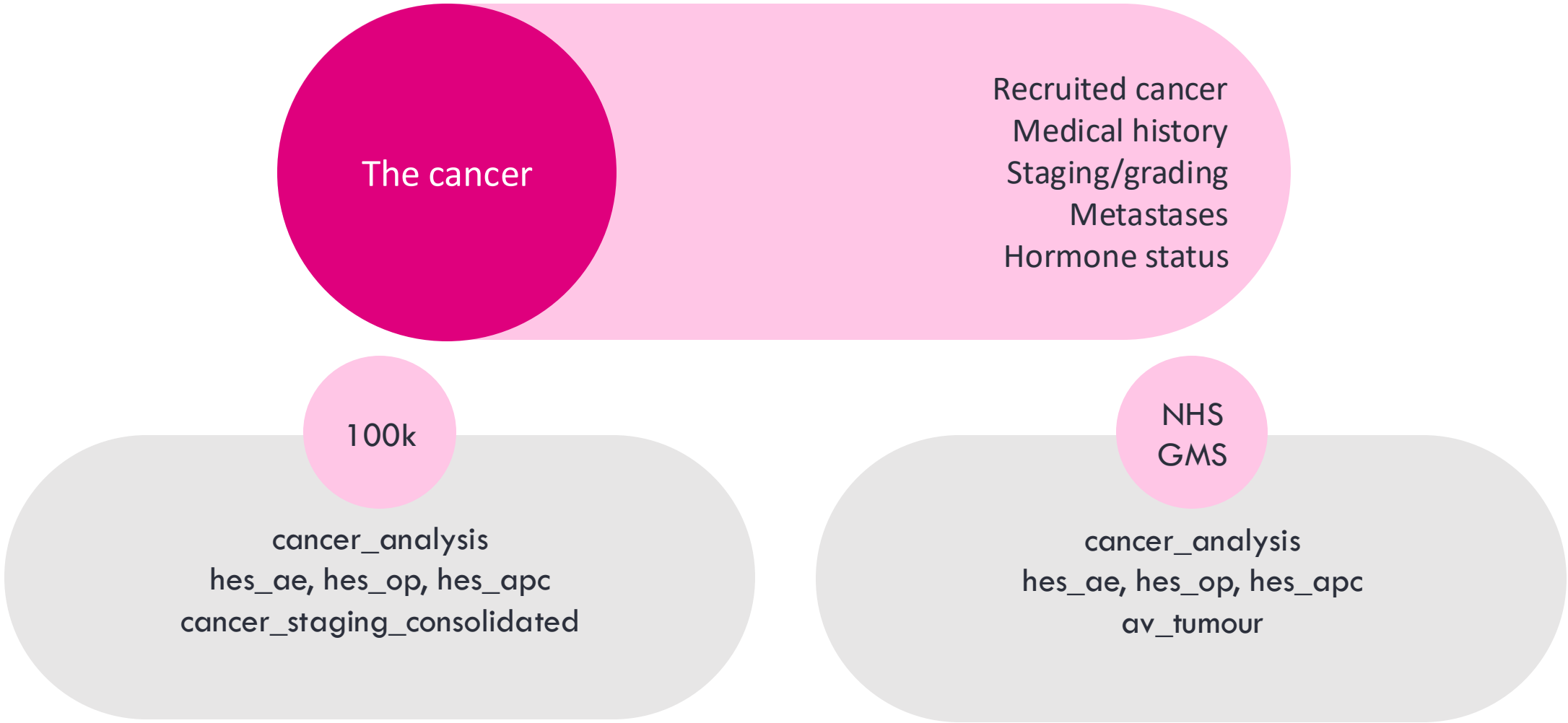
https://re-docs.genomicsengland.co.uk/pxa/

Participant Explorer demo

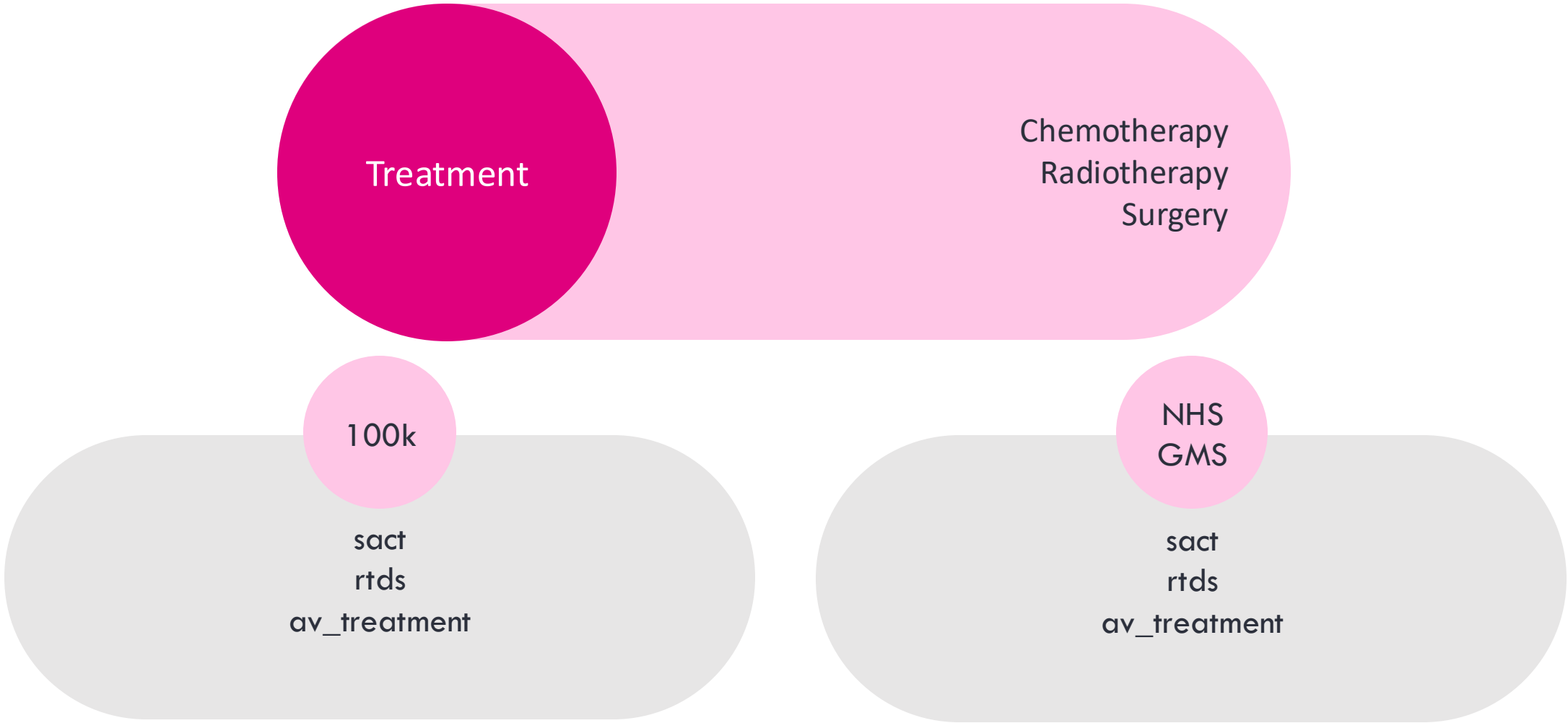

Camtasia File Edit Modify Text View Export Window Help
In 25m: Double Helix user e...
Wed 2 Apr 10:35
Amazon WorkSpaces
Computer
Document Viewer
Open Targets
Visual Studio Code
Participant Explorer
eperry's Home
LibreOffice 7.6
RE Messages
Trash
Link to emily
Emacs
Research Registry
R
Old Firefox Data
Ensembl
Dremio
Text Editor
Airlock
GVim
Research Environment Documentation
Welcome Pack
CloudOS Academic
IGV Browser
CloudOS Discovery Forum
IVA
RStudio
Firefox
CloudOS Internal
Labkey
Terminal Emulator
Git GUI
Desktop.Rproj
Panel App
Genomics england
Applications Places System
[eperry@a-fopkxs25p4f...
[cohort_building_trainin...
[cancer_cohorts_2025]
[~/pgen_int_work/BRS/...
[Terminal]
[Mozilla Firefox]
Wed Apr 2, 10:35

4. Tables for cohort building in cancer

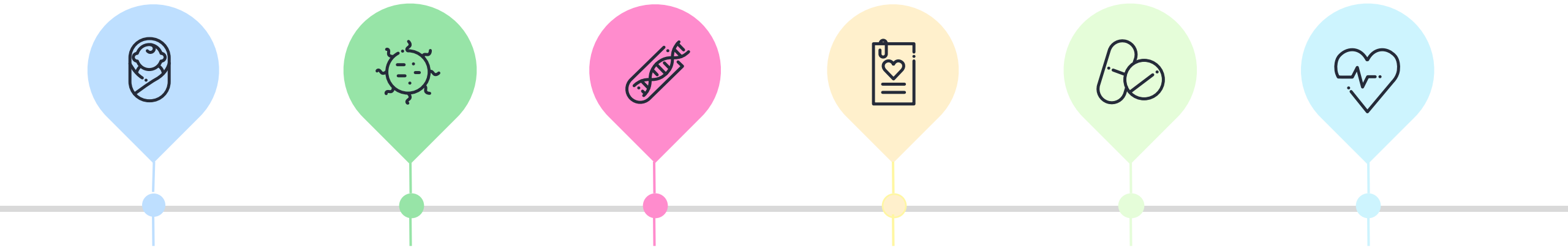
Cancer cohort parameters



Cancer cohort parameters



Derive age from dates



Year of birth

Diagnosis

Sampling

Phenotyping

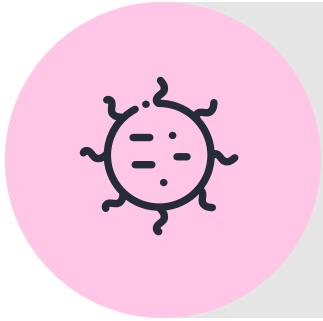
Treatment

Death

Keys

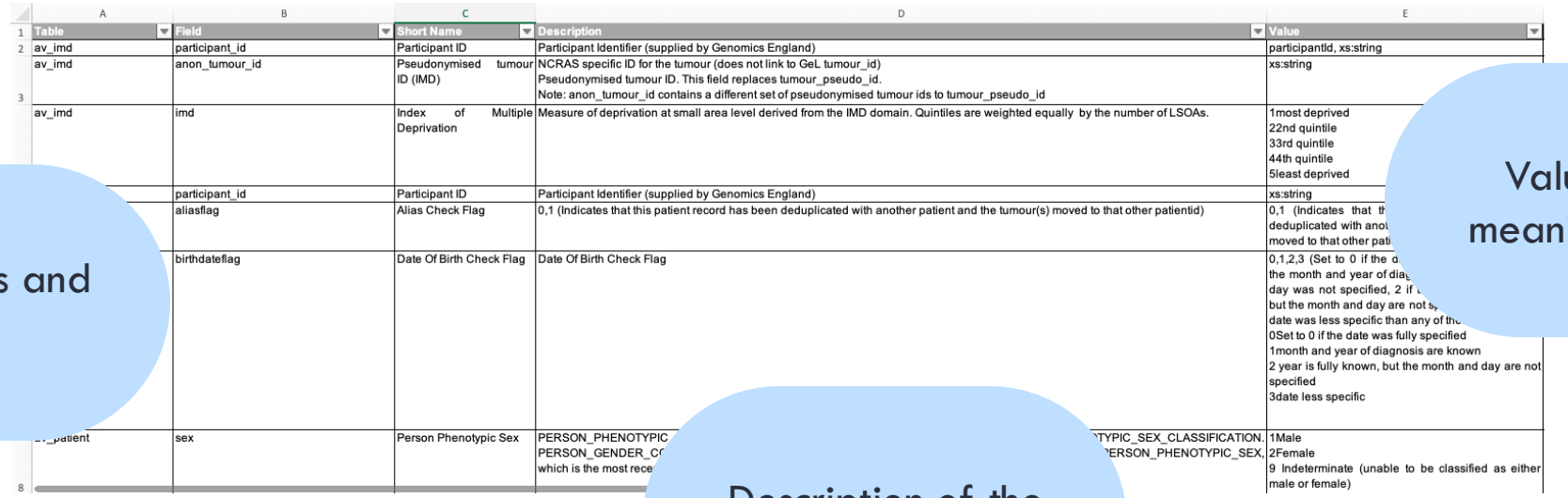
participant_id

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 diag day was not specified, 2 if but the month and day are not s date was less specific than any of the 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	_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)

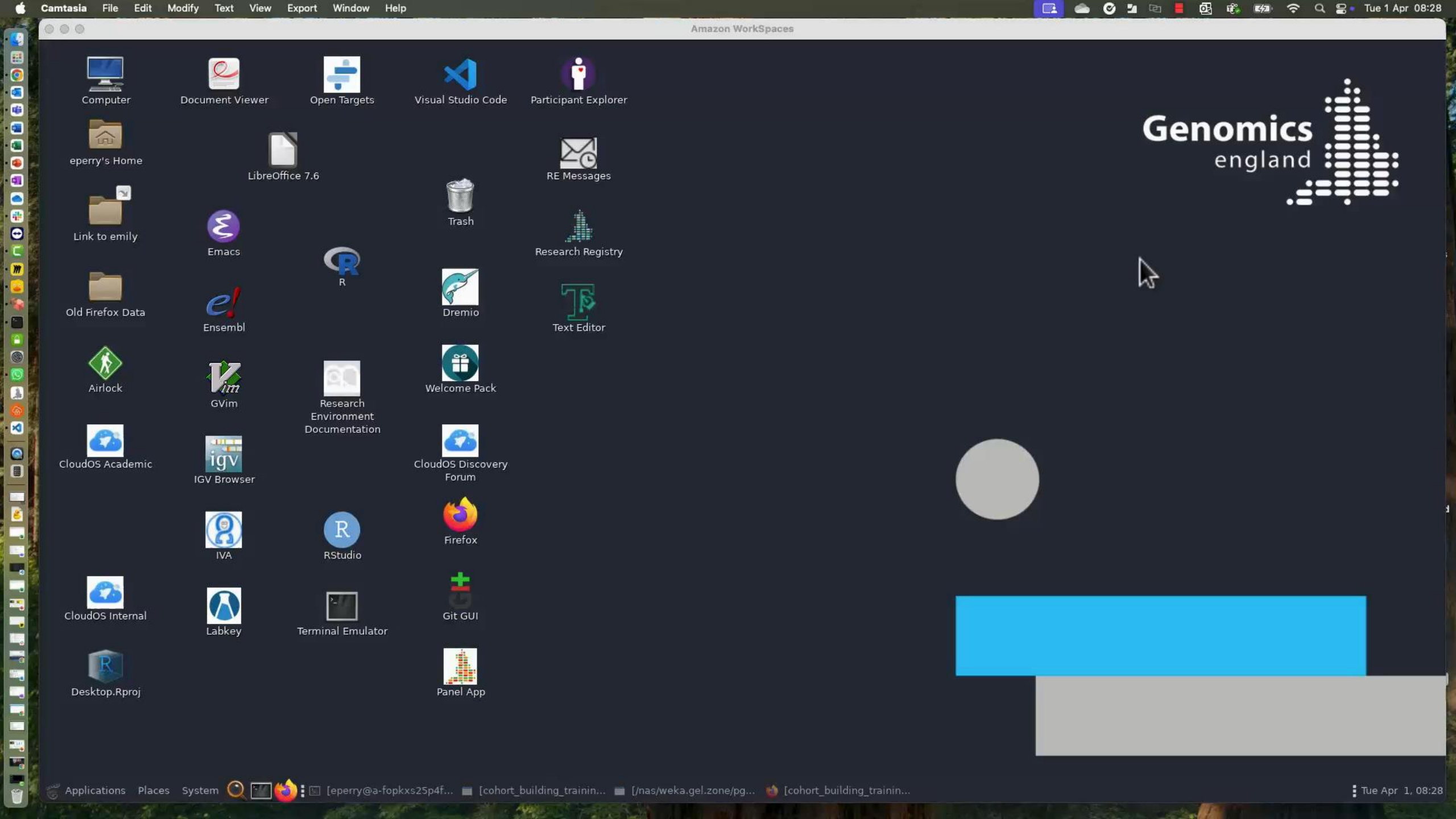
Lists of tables and columns

Value type or meaning of codes

Description of the data

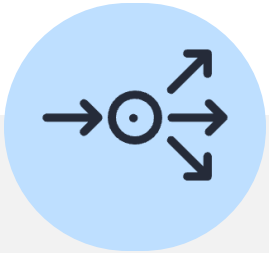
https://re-docs.genomicsengland.co.uk/100k_data_dictionary https://re-docs.genomicsengland.co.uk/GMS_data_dictionary

Tables demo

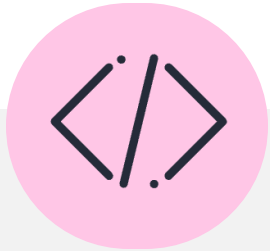

Camtasia File Edit Modify Text View Export Window Help
Tue 1 Apr 08:28
Amazon WorkSpaces
Computer
Document Viewer
Open Targets
Visual Studio Code
Participant Explorer
eperry's Home
LibreOffice 7.6
RE Messages
Trash
Link to emily
Emacs
Research Registry
R
Old Firefox Data
Ensembl
Dremio
Text Editor
Airlock
GVim
Research Environment Documentation
Welcome Pack
CloudOS Academic
IGV Browser
CloudOS Discovery Forum
IVA
RStudio
Firefox
CloudOS Internal
Labkey
Terminal Emulator
Git GUI
Desktop.Rproj
Panel App
Genomics england
Applications Places System
[eperry@a-fopkxs25p4f...
[cohort_building_trainin...
[/nas/weka.gel.zone/pg...
[cohort_building_trainin...
Tue Apr 1, 08:28

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_2025

Programming demo

mortality: /main-prog | pgen_int_work/BRS/e | cohort_building_train

localhost:8387/notebooks/pgen_int_work/BRS/emily/cancer_co

jupyter cohort_building_training (autosaved) Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3 (ipykernel)

Run Markdown

Import Python modules you need

In [1]:

```
import numpy as np
import functools
import labkey
import pandas as pd
```

Helper function to access the LabKey API with Python

The following Python function accesses the LabKey API, executes an SQL query to the specified database and converts it to a dataframe using pandas. We recommend putting this at the beginning of your scripts/notebooks and invoking it every time you want to access the database.

In [2]:

```
def labkey_to_df(sql_query, database, maxrows):
  """generate an pandas dataframe from labkey sql query

  Args:
    sql_query (str): an sql query as string.
    dr (str): GEL datarelease version
    """
  ver = labkey.__version__

  if ver == '1.2.0' or ver == '1.4.0' or ver == '1.4.1':
    server_context = labkey.utils.create_server_context(
      domain= "labkey-embassy.gel.zone",
      container_path = database,
      context_path = "labkey",
      use_ssl = True
    )

    results =  labkey.query.execute_sql(
      server_context,
      schema_name="lists",
      sql=sql_query,
      max_rows=maxrows
    )
```

~/pgen_int_work/BRS/emily/cancer_cohorts_2025/cohort_building_training/Notebook.nb.html

Notebook.nb.html | Open in Browser | Find | Publish

Cancer cohort building in R

Code

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

Import R libraries you need

Hide

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.2     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.2     ✔ tibble    3.2.1
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.1     ── Conflicts ────────────────────────────── tidyverse_conflic
```

6. Getting genomic filepaths for your cohort

Genomic files available

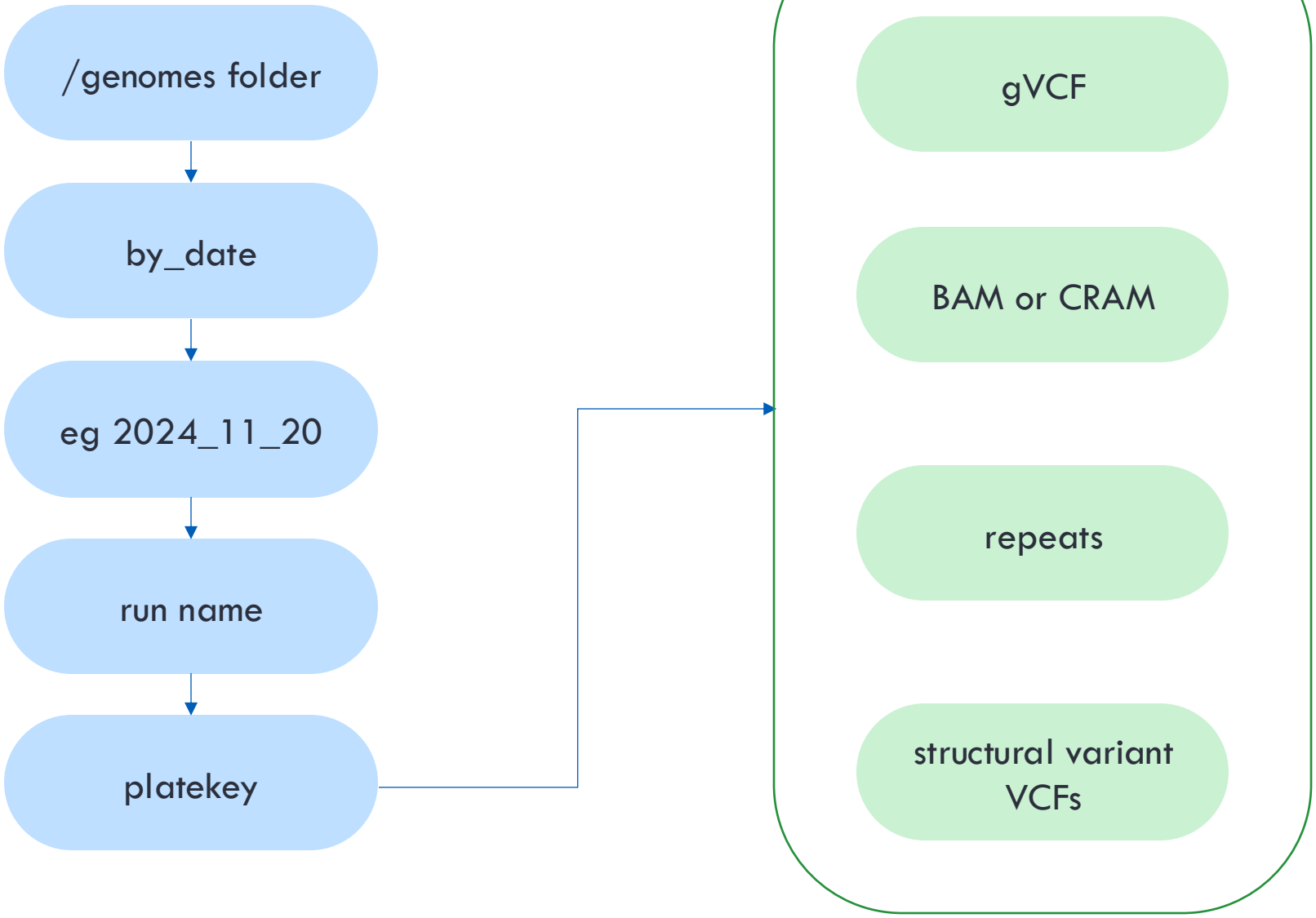
Germline

Platekey
BAM
gVCF
SV VCF

Tumour

Platekey
BAM
gVCF
SV VCF

Genomes location

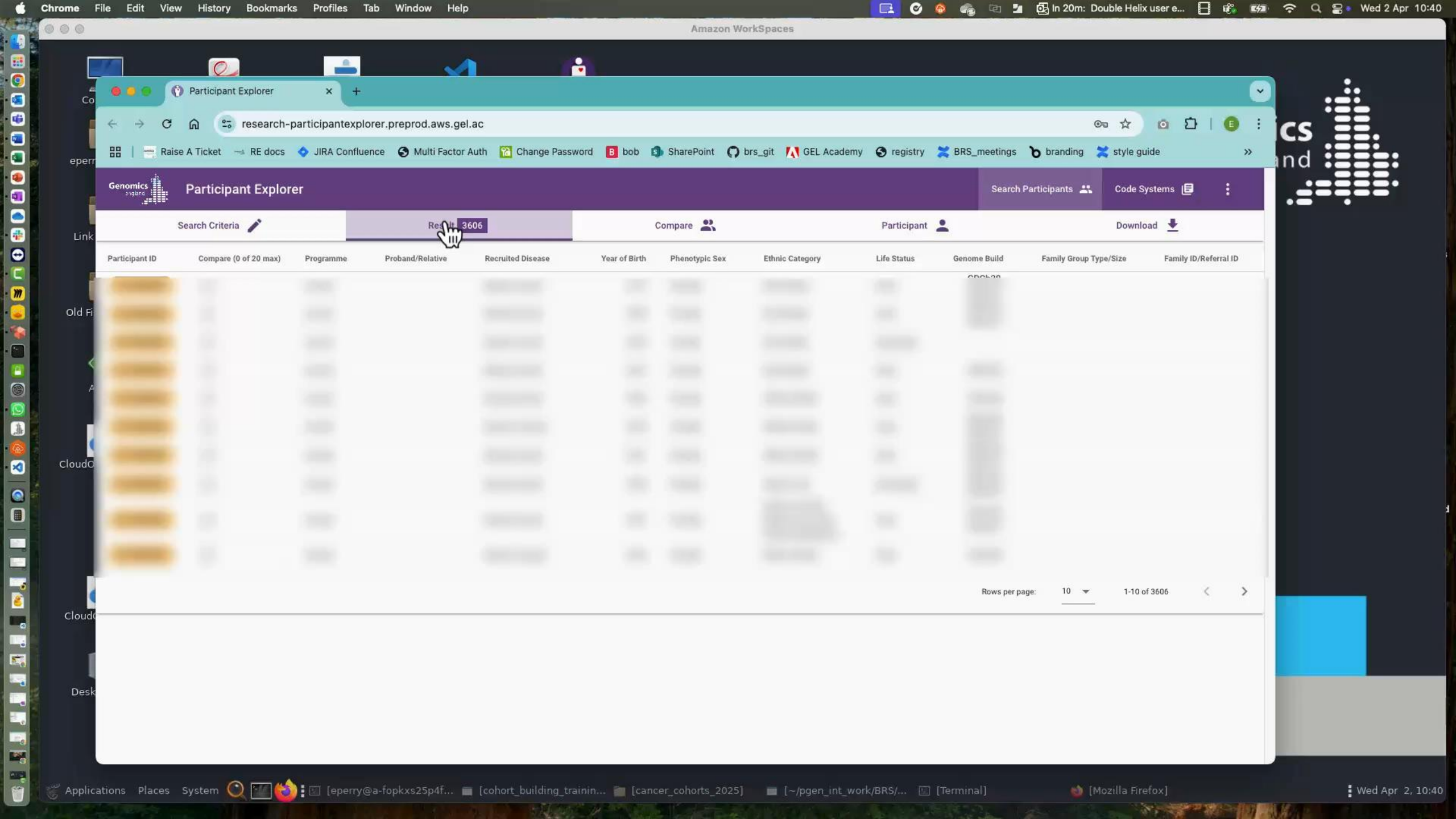


https://re-docs.genomicsengland.co.uk/genomic_data/

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


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Wed 2 Apr 10:40
Amazon WorkSpaces
Participant Explorer
research-participantexplorer.preprod.aws.gel.ac
Raise A Ticket
RE docs
JIRA Confluence
Multi Factor Auth
Change Password
bob
SharePoint
brs_git
GEL Academy
registry
BRS_meetings
branding
style guide
Genomics England
Participant Explorer
Search Participants
Code Systems
Search Criteria
Result 3606
Compare
Participant
Download
Participant ID
Compare (0 of 20 max)
Programme
Proband/Relative
Recruited Disease
Year of Birth
Phenotypic Sex
Ethnic Category
Life Status
Genome Build
Family Group Type/Size
Family ID/Referral ID
Rows per page:
10
1-10 of 3606
Applications
Places
System
Wed Apr 2, 10:40

7. Using your cohort with aggregate VCFs

Aggregate VCFs

Combined gVCFs

Quality metrics

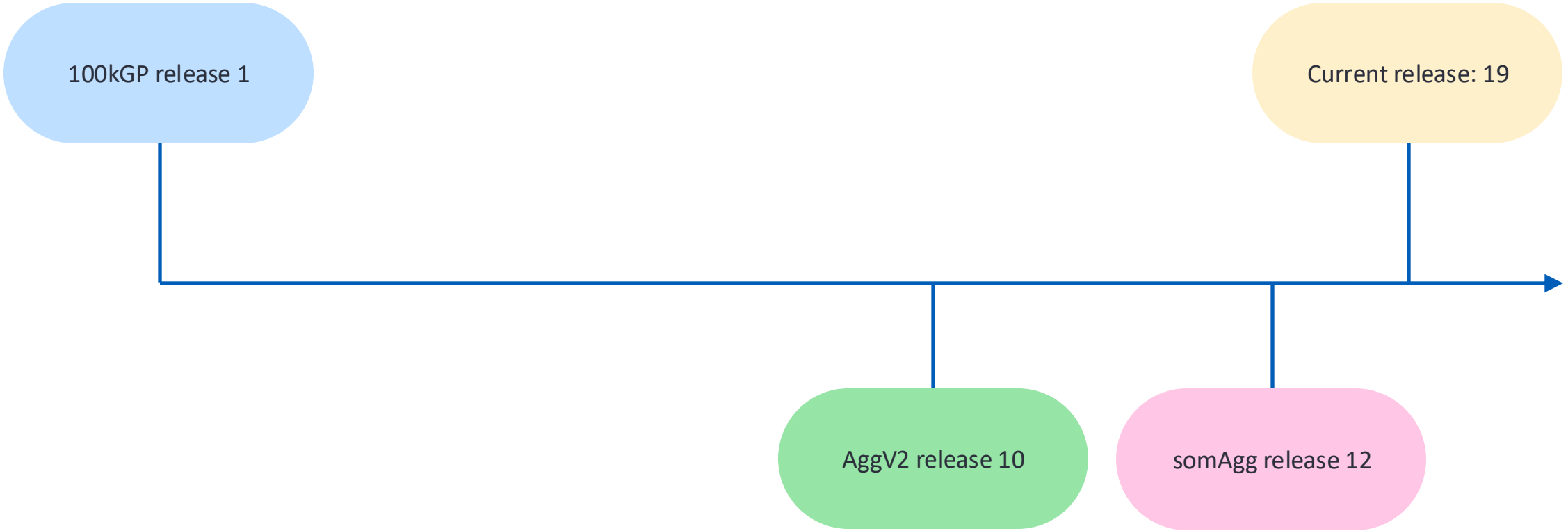
Functional annotation files

Genotype files – platekeys

1371 chunks

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

Aggregate VCFs



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/

Aggregate demo

Camtasia File Edit Modify Text View Export Window Help

Tue 1 Apr 13:26

Amazon WorkSpaces

cancer_analysis: /nh | pgen_int_work/BRS/e | cohort_building_train

localhost:8387/notebooks/pgen_int_work/BRS/emily/cancer_co

jupyter cohort_building_training (autosaved)

Logout

File Edit View Insert Cell Kernel Help Trusted Python 3 (ipykernel)

Run Code

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

In [41]:
```
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 100kGP cohort. Let's start by creating a list of platekeys to use for our query.

In [42]:
```
platekeys = ','.join(path_query['germline_sample_platekey'])
platekeys
```

~/pgen_int_work/BRS/emily/cancer_cohorts_2025/cohort_building_training/Notebook.nb.html

Notebook.nb.html | Open in Browser | Find | Publish

```
        gms_filetype,
        ", germline_sample_platekey
    FROM cancer_analysis
    WHERE participant_id IN ('", paste(gms_list, collapse = "', '"), "')", sep="")

gms_path_query <- labkey_to_df(gms_path_sql, gms_version, 100000)
gms_path_query
```

participant_id
<chr>

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

Hide

```
write.csv(path_query,file='r_100k_paths.csv', row.names=FALSE)
write.csv(gms_path_query,file='r_gms_paths.csv', row.names=FALSE)
```

Working with aggregate VCFs

Applications Places System [eperry@a-fopkxs25p4f... [cohort_building_trainin... cohort_building_trainin... [cohort_building_training] ~/pgen_int_work/BRS/e...

Tue Apr 1, 13:26

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

In-person training

30/6 For non-coders

1/7 For coders

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

Training sessions

3rd Tuesday every month

Introduction to the RE

20/5 22/7 19/8

16/9 21/10 18/11

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

Training sessions

10/6	Building rare disease cohorts with matching controls
8/7	Finding participants based on genotypes
9/9	Getting medical records for participants
14/10	What tools and workflows should I use to fulfil an overall goal?
11/11	Using GEL data for publications and reports
9/12	Running workflows on the HPC and Cloud

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

Feedback

Thank you

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