

## Building a cohort based on phenotypes

**Emily Perry** 

Research Engagement Manager

24<sup>th</sup> May 2022

#### 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
- 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)
- We will record this training and distribute the censored video afterwards

#### Questions



Your microphones are all 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.









**Christian Bouwens** Bioinformatician -Research Services

Ronnie Rodrigues Pereira Bioinformatician -Research Services





Roel Bevers Senior Bioinformatician -Research Services



3

4

5

Introduction and admin



Parameters and considerations for building a cohort

- Point-and-click cohort building with Participant Explorer
- Labkey tables for cohort building
- Covariates in cohort building
- 6

7

8

- Using the Labkey API in Python and R to build cohorts
- Fetching data for downstream analysis
- Getting help and questions

#### Materials

- Slides and video will be sent out to you after the session
- Scripts available in /gel\_data\_resources/example \_scripts/workshop\_scripts/coh ort\_building\_20220524

20220524 - File Manager	-		×
elp			
<pre>cohort_building_20220524 cohort_building_training</pre>			>
cohort building t cohort building t			
raining raining.ipynb			
2 itoms (27.2 kP) Error space: 2.7 PP			
	20220524 - File Manager elp Cohort_building_20220524 Cohort_building_training Cohort_building_t cohort_building_t raining raining.ipynb	20220524 - File Manager	20220524 - File Manager – – – – – – – – – – – – – – – – – – –

## 2. Parameters and considerations for building a cohort

#### Phenotype data to filter by



- Disease type
  - Recruited disease
  - Diagnosis codes in health records
- Staging
- Treatment

#### Rare disease

- Recruited disease
- HPO terms
- Solved? Alive?

#### Common disease

 Rare disease relatives and cancer participants
 Diagnosis codes in health records

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

#### Participant Explorer

- Point-and-click tool
- Ontology-aware
- Filter by phenotypes
- Combine multiple filters



### Demo: Participant Explorer

https://research-help.genomicsengland.co.uk/display/GERE/Participant+Explorer







۵	Particip	ant Explorer — Mo	zilla Firefox		, T			_	٥	×
() R	estore Sess	sion ×	<pre>c pgen_int_work/BRS/smacx</pre>	릗 pcdh19_repo	rt - Jupyter 🔯 agg	$gregate_gvcf_sample_{\times}$	🔕 rare_diseases_participan ×	Participant Explorer	×	) +
$\leftarrow$	$\rightarrow$ G		O A https://prod.termino	logy-service.aws. <b>g</b> e	el.ac			☆	${igsidential}$	=
Geno	omics	Participant Ex	kplorer				Search Participa	nts 🚢 Code Systems 🖻	:	
	Sear	rch Criteria 🧪	Result		Compare		Participant	Download 👤		
	÷	Search By Clinic	cal Concept						a <sup>K</sup>	
	•		Match Any / Match None	A 11	Filter By Code Set	Search Concepts				
	Ĥ		Match Any	All	All	Type at least 3 characte	ers to search	•		
	¢			Condition Observation Procedure		Inc	clude descendant concepts			
		Search by Partie	cipant Details	Body Site					и <sup>л</sup>	
		Search								
	pplications	ۏ Participant E	xplorer — 💽 [Terminal]							09:3

🔞 Participa	ant Explorer — Mozi	illa Firefox		× ·				-	đ	×
(i) Restore Sessi	ion ×	pgen_int_work/BRS/smac×	🖉 pcdh19_report	- Jupyter 🗠 🚺 aggregate_g	gvcf_sample_ $\times$	🚺 rare_diseas	ses_participan ×	Participant Explorer	×	+
$\leftarrow \  \   \rightarrow \  \   G$		O A https://prod.termind	ology-service.aws. <b>gel.</b>	ac				☆	${igsidential}$	=
Genomics	Participant Exp	blorer					Search Participants	🙁 Code Systems 🗐	:	
Sear	ch Criteria 🧪	Result	=	Compare		Participant		Download 👤		
•	Search By Clinica	l Concept							л <sup>к</sup>	
			Filter Dr. Oliviani Over							
Â		Match Any	All	All	least 3 characters	s to search		•		
¢			e mapped concepts 🕐	SNOMED CT Human Phenotype Ontology	Inclu	ude descendant	concepts			
				Recruited Diseases					_	
	Search by Partici	pant Details		ICD-10					к <sup>и</sup>	
				OPCS-4						
	Search			ICD-0-3						
Applications	🍪 Participant Exp	plorer — 💽 [Terminal]								09:40

6	Participant E	xplorer —	Mozilla Fire	efox						Ð	×
(i) Resto	ore Session	>	c pg	en_int_work/BRS/smac×	pcdh19_report	- Jupyter 🔯 agg	gregate_gvcf_sample_ $\times$	▲ rare_diseases_participani×	Participant Explorer	×	) +
$\leftarrow \rightarrow$	С		0	https://prod.termino	logy-service.aws. <b>gel.</b>	ac				${igsidential}$	≡
Genomics england	Par	ticipant	Explore	r				Search Participants	s 🐣 Code Systems 🖻	:	
	Search Cri	iteria 🦯		Result	=	Compare		Participant 💄	Download 💻		
	Se	earch By Cl	inical Conc	ept						л <sup>к</sup>	
				Match Any / Match None Match Any	Filter By Clinical Conc	Filter By Code Set	Search Concepts				
	<u></u>										
	6			Include	mapped concepts ()		65 matches found for "colore	Conomics England Canoer			
_							Colorectal	OPCS Z294	Туре		
	5	arah hu Dr	urticipant D	ataila			Colorectal polyposis	HP0 HP:0200063		2	
	Se	earch by Pa	irticipant D	etans			Proctocolitis	SNOMED 418130002		¥″	
							Porter colored urine	SNOMED 720002008			
		Sear	ch				Skin-colored papule	HP0 HP:0025512			
							Coloproctostomy	SNOMED 70243005			
							Wears brightly colored clo	othes SNOMED 225498000			
							Things appear vividly colo	ored SNOMED 247717002			
E Appli	cations 🚯 P	Participan	t Explorer	🔄 [Terminal]							09:4

🔞 Particip	ant Explorer — Mozilla Fire	fox						-	đ	×
(i) Restore Sess	sion × Ċ pg	en_int_work/BRS/smac×	pcdh19_report	- Jupyter 🔯 agg	regate_gvcf_sample_ $\times$	🚺 rare_disea	ses_participani×	Participant Explorer	×	+
$\leftarrow \ \rightarrow \ G$	0	https://prod.terminol	ogy-service.aws. <b>gel.</b>	ac				<del>ت</del>	$\bigtriangledown$	Ξ
Genomics	Participant Explore						Search Participants	🐣 Code Systems 🖻	÷	
Sea	rch Criteria 🧪	Result		Compare		Participant		Download 👤		
*	Search By Clinical Conce	ept							л <sup>к</sup>	
2. A		Match Any / Match None Match Any	Genomics England Ca	ancer Type   Colorectal 🛛 🗙	Descendant concepts i	ncluded		î		
¢			Filter By Clinical Conc	Filter By Code Set	Search Concepts Type at least 3 character	rs to search		•		
		Include t	mapped concepts ①			lude descendant	concepts	$\oplus$		
	Search by Participant De	etails							κ <sup>π</sup>	
	Search									
Applications	🔞 Participant Explorer	— 📧 [Terminal]								09:42

iii Participa	nt Explorer –	- Mozilla Fire	ox								٥	×
(i) Restore Sessio	on ×	🔵 pgen_in	t_work/BRS/S×	/ pcdh19_repo	rt - Jupy×	🐼 cancer_analysis: /r	ma×	🚺 sact: /main-program×	M rare_diseases_partic×	😗 Participant Exp	plorer $\times$	) +
$\leftarrow \rightarrow G$		08	https://prod.t	erminology-servi	ce.aws. <b>ge</b> l	l.ac				☆	${igsidential}$	=
Genomics	Participan	t Explorer	X.						Search Participants 🚢	Code Systems ᄐ	:	
Searc	ch Criteria 🧪			Result 📃		Compare		Particip	pant 🚨	Download 💻		
<b>±</b>	Search By C	Clinical Conce	pt								7 <sup>K</sup>	
•			Match Any / Match N	one								
â			Match Any	Genomi	cs England C	cancer Type   Colorectal	Ge	enomics England Cancer Type	Breast X	Ĩ		
¢				Descent	uant concept							
				Filter By Cl	inical Conc	Filter By Code Set	Search Type	n Concepts at least 3 characters to searc	sh 🔹	~		
				Include mapped co	oncepts ()			Include desc	endant concepts	$\oplus$		
	Search by P	Participant De	tails								и <sup>Я</sup>	
Applications	᠔ Participaı	nt Explorer	— 🔄 (Termi	nal]								09:55

🔞 Partic	ipant Explorer — Moz	zilla Firefox					-	ð	×
(i) Restore Se	ession × 😋 p	gen_int_work/BRS/s×	pcdh19_report - Jupy×	🔕 cancer_analysis: /m	× 🔕 sact: /main-program×	▲ rare_diseases_partic×	Participant Expl	orer ×	+
$\leftarrow \rightarrow \mathbf{G}$	;	O A https://prod.t	terminology-service.aws. <b>ge</b>	l.ac			Participant Expl ☆	lorer J	≡
Genomics	Participant Ex	plorer				Search Participants 🙁	Code Systems 🗗	:	
Se	earch Criteria 🧪		Result	Compare	Particip	ant 💄	Download 星		
±	Search By Clinica	al Concept						л <sup>к</sup>	
2			_						
â		Match Any	Genomics England C	Cancer Type   Colorectal	Genomics England Cancer Type   Brea	st Descendant concepts incl	uded		
¢	And / Or And		Ione Filter By Clinical Conc All	Filter By Code Set All	Search Concepts Type at least 3 characters to searc	h	×		
		0	Include mapped concepts 🕐		Include desce	endant concepts			
	Search by Partic	ipant Details						κ <sup>π</sup>	
	Search								
Application	ns 🄞 Participant Ex	kplorer — 💽 [Termi	inal]					(	09:5

🔞 Particip	pant Explorer — Mozilla	a Firefox		v			<u> </u>	۵ ×	×
(i) Restore Sess	sion × 📿 pge	en_int_work/BRS/SX	cdh19_report - Jupy×	🔕 cancer_analysis: /m	× 🚺 sact: /main-program ×	rare_diseases_partic×	() Participant Expl	orer ×	+
$\leftarrow \ \rightarrow \ G$	1	C A https://prod.termin	ology-service.aws. <b>ge</b> l	l.ac			☆		=
Genomics	Participant Expl	orer				Search Participants 🚢	Code Systems 🔳	:	
Sea	arch Criteria 🧪	Result	=	Compare	Partic	ipant 🔔	Download 🛓		
<b>.</b>	Search By Clinical	Concept						я <sup>к</sup>	
		Match Any	Genomics England C	ancer Type   Colorectal	Genomics England Cancer Type   Bre	east Descendant concepts inclu	ded		
•	And / Or	Match Any / Match None	Filter By Clinical Conc	Filter By Code Set	Search Concepts				
¢	And	Match Any	All	All	tonsille	•	×		
	-				28 matches found for "tonsille" Incl	ude all < Page 1 / 3 💙			
			e mapped concepts ()		Tonsillectomy	SNOMED 173422009			
					Dissection tonsillectomy	SNOMED 173424005			
	Search by Participa	int Details			Guillotine tonsillectomy	SNOMED 173425006		л	
					Histology tonsillectomy	SNOMED 276748006			
		-			Radiofrequency tonsillectomy	SNOMED 854041000000104			
	Search				Remnant tonsillectomy	SNOMED 173428008			
					Tonsillectomy planned	SNOMED 183979001			
Applications	s 🥹 Participant Expl	orer — 📧 [Terminal]						09	9:5

۲	Participa	ant Explorer — Mozill	la Firefox					_	đ	×
() I	Restore Sessi	ion × 😄 pg	en_int_work/BRS/S× 🗧 pc	dh19_report - Jupy×	🔕 cancer_analysis: /ma	sact: /main-program×	🔕 rare_diseases_partic×	() Participant Explo	rer ×	+
$\leftarrow$	$\rightarrow$ G		O A https://prod.terming	logy-service.aws. <b>gel.a</b>	ac				$\bigtriangledown$	=
	Sear	rch Criteria 🧪	Result	=	Compare	Particip	ant 🔔	Download 👤		
	*	Search By Clinical	Concept						я <sup>к</sup>	
	•)									
	â		Match Any	Genomics England Car	ncer Type   Colorectal G	enomics England Cancer Type   Brea	st Descendant concepts incl	uded		
	¢	And / Or And	Match Any / Match None Match Any	SNOMED   173422009:	Tonsillectomy 🗴 Des	cendant concepts included		Î		
				Filter By Clinical Conc All	Filter By Code Set	earch Concepts ype at least 3 characters to searc	h 👻	~		
								÷		
			Include	mapped concepts ()			endant concepts			
		Search by Particip	ant Details						к <sup>л</sup>	
		Search								
	Applications	🕹 Participant Exp	lorer — 🔄 [Terminal]							10:00

(1)	Participa	ant Explorer — Mozilla	Firefox		×.			-	٥	×
() I	Restore Sessi	ion × Ċ pge	en_int_work/BRS/S× 🧧 p	ocdh19_report - Jupy×	Cancer_analysis: /ma	🔇 🔕 sact: /main-program×	▲ rare_diseases_partic×	() Participant Explo	rer ×	+
$\leftarrow$	$\rightarrow$ C	(	C A https://prod.termi	nology-service.aws. <b>gel.</b>	ac			٢	$\bigtriangledown$	Ξ
	Sear	ch Criteria 🧪	Resu	lt 🔳	Compare 🏔	Particip	pant 🚨	Download 👤		
	<b>±</b>	Search By Clinical C	Concept						א <sup>⊭</sup>	
	-		Match Any	Genomics England Ca	ancer Type   Colorectal G	enomics England Cancer Type   Brea	ast Descendant concepts inc	luded		
				Mapped concepts for	ind: 0					
	¢	And / Or	Match Any / Match None							
		And	Match Any	SNOMED   173422009	9: Tonsillectomy 🗙 Desc	cendant concepts included Ma	pped concepts found: 9	Ī		
				<> OPCS   F346: Ex	cision of lingual tonsil	···› OPCS   F342: Bilateral guillotine	tonsillectomy			
				<> OPCS   F343: Bi	lateral laser tonsillectomy	OPCS   F344: Bilateral excisi	on of tonsil NEC			
				<b>&lt;&gt;</b> OPCS   F345: Ex	cision of remnant of tonsil		ed excision of tonsil			
					nspecified excision of tonsil	⟨··⟩ OPCS   F347: Bilateral cobl	lation tonsillectomy			
				<ul> <li>&lt;&gt; OPCS   F341: Bi</li> </ul>	lateral dissection tonsillectom	у				
				Filter By Clinical Conc	Filter By Code Set	earch Concepts	sh 👻	,		
				All	AII	ype acteast 5 characters to sealt	211	~		
	Applications	💩 Participant Explo	orer — 🔄 [Terminal]							10:0



O Restore Session × ● pren_int_work/8 × ● pcdh19_report · × ③ cancer_staging × ③ sact: /main-prog × ④ rare_diseases_pi × ④ mortality: /main × ④ Participant Expl × 2               C          O A          A https://prod.terminology-service.aws.gel.ac         O Search by Participant Details              Programme         Vear of Birth Range         1900 ④         O Search by details reb         Phenotypic Sex         Life Status         Phenotypic Sex         Life Status         Programme specific search fields         Proband/Relative         Affection Status         Family Group Type         Fam	🔞 Participa	ant Explorer — Mozilla Firefox			×			- 0	×
C        C	(i) Restore Sessi	ion × $\bigcirc$ pgen_int_work/B ×	pcdh19_report - ×	Cancer_staging_	🔕 sact: /main-prog ×	▲ rare_diseases_pax	🔕 mortality: /main- ×	🕐 Participant Explox	+
Sarch by Patiopant Details           Year of Birth Range         Programme         100	$\leftarrow \  \   \rightarrow \  \   G$	0 8 ~ h	tps://prod.terminology-se	rvice.aws.gel.ac			☆		≡
Search by Participant Details     Year of Birth Range     Programme     1900        Phenotypic Sex     Life Status        Rare Disease Programme specific search fields     Proband/Relative     Affection Status     Family Group Type     Family Group Type     Family Case Solved									
Programme     Phenotypic Sex           Life Status        Ethnic Category              Proband/Relative     Affection Status     Family Group Type     Family Case Solved		Search by Participant Details						۶ <sup>۲</sup>	
Phenotypic Sex     Life Status        Ethnic Category     Rare Disease Programme specific search fields     Proband/Relative     Affection Status     Family Group Type     Family Case Solved		Programme		Year of Birth Range		● 2020 ≎	① Search	by details help	
Proband/Relative       Affection Status       Family Group Type       Family Case Solved       Family Case Solved		Phenotypic Sex 🔻	Life Status	*	Ethnic Category	••	enome	e Build 👻	
Proband/Relative  Affection Status Family Group Type Family Group Type Family Case Solved Family Case Solved				Rare Disease	Programme specific search fie	elds			
		Proband/Relative 🔻	Affection St	atus 🔻	Family Group Typ	e 💌	Family C	Case Solved 🔻	
Search		Search							

🝅 Participant Explorer — Mozilla Firefox	Y			- 0	×
(i) Restore Session × 📿 pgen_int_work/B	< 🧧 pcdh19_report - × 🚺 cancer_staging_ × 🚺 sac	:t: /main-prog ×	🚺 mortality: /main- × 🕅 Parti	icipant Explox	) +
$\leftarrow \rightarrow \mathbf{G}$ $\bigcirc \mathbf{P} \diamond$	https://prod.terminology-service.aws. <b>gel.ac</b>		☆	${igardown}$	=
Search by Participant Details				۶ <sup>۲</sup>	
	Year of Birth Range				
Programme 🔻	1900 🗢 🍙	● 2020 ♥	① Search by details here	ielp	
Phenotypic Sex 🔻	Life Status 💌	Other Ethnic Groups: Any other ethnic gr	roup Genome Build	•	
		Other Ethnic Groups: Chinese	_		
	Rare Disease Programn	N White: Any other White background			
Proband/Relative	Affection Status	White: British	Family Case Solver	ed 🔻	
		White: Irish			
Search		empty			
Applications  Participant Explorer —	🖻 [Terminal]				11:1

🍅 Participa	ant Explorer — Mozilla Fir	refox			V				-	٥	×
(i) Restore Sess	ion × Ċ pgen_	int_work/BRS/S×	apcdh19_rep	ort - Jupy × 🛛 🛛 ca	ancer_analysis: /ma×	🐼 sact: /m	ain-program × 🛛 🚺	rare_diseases_partic×	😗 Participant Exp	olorer ×	] +
$\leftarrow \  \  \rightarrow \  \  \mathbf{G}$	0	A https://proc	l.terminology-serv	vice.aws.gel.ac						$\bigtriangledown$	
Genomics	Participant Explore	er						Search Participants 🚢	Code Systems 트		
Sear	rch Criteria 🧨		Result 32		Compare		Participant		Download 👤		
Participant ID	Compare (0 of 20 max)	Programme	Proband/Relative	Recruited Disease	Year of Birth	Phenotypic Sex	Ethnic Category	Consent Form Li	fe Status Genome B	uild	Famil
-		Cancer		Breast, Ductal		Female	White: British	Adult C1	GRCh	138	
-		Cancer		Colorectal, Adenocarcinoma		Male	White: British	Adult C1	GRCh	138	
		Cancer		Colorectal, Unknown		Male	White: British	Adult C1 De	eceased GRCh	138	
-		Cancer		Breast, Unknown		Female	White: British	Adult C1	GRCh	138	
-		Cancer		Breast, Unknown		Female	Not Stated	Adult C1			
_		Cancer		Colorectal, Adenocarcinoma		Female	White: British	Adult C1	GRCh	138	
		Cancer		Breast, Lobular		Female	White: British	Unknown			
_		Cancer		Breast, Ductal		Female	Not Known	Adult C1	GRC	138	
-		Cancer		Colorectal, Adenocarcinoma		Male	White: British	Adult C1			
-		Cancer		Breast, Ductal		Female	Other Ethnic Groups: Any other ethnic group	Adult C1	GRCł	138	
								Down not not not 10	1 10 -6 22	,	
Applications	ۏ Participant Explore	er — 📧 [Terr	minal]					ROWS DEL DAGE. 10	- 1-10 of 37	(	10:03

🔞 Participant Explorer — Mo	zilla Firefox						-		×
(i) Restore Session $\times$ $\bigcirc$ (	pgen_int_work/BRS/EX	🧧 pcdh19_report - Jupy×	Cancer_analys	is: /max	🔕 sact: /main-program×	🔕 rare_diseases_partic×	Participant Explor	er ×	+
$\leftarrow \rightarrow G$	O A https://prod.te	erminology-service.aws. <b>gel.</b>	ac						=
Genomics england 	cplorer					Search Participants 🚢	Code Systems   🖻	:	
Search Criteria 🧨		Result 32	Compare	-	Participan	±	Download 👤		
Available Columns				Selected Colu	umns				
Participant 🖬 Ethnic Category Family ID Recruited Disease 🔹	Family Group Type	Consent Form		Participant Participan Phenotypic	t ID Programme Proba	nd/Relative Year of Birth	Date of Death		
Cancer Disease Type Canc Rare Disease Sub Group S Rare Diseases Family Case Repor	t	re Disease Group		Notes: • The resul • The resul example, • Rare Dise	It contains unique rows only. It may contain multiple rows per pa , multiple values for Genome Build a eases Family Case Report, Conditio	rticipant, when columns with mu are rendered as separate rows. n/Observation/Procedure and G	Iltiple values are selected. F enome Sequence columns (	or cannot	
Interpretation Request ID F	Family Case Solved Add	tional Comments		be combi • When Par each uniq	ined in the same result. rticipant ID is not selected, a colum que row.	nn is included with the count of p	articipants in the search re	sult for	
Clinical Concept Code Syst	tem Code Descript	ion Source Code		Download	<b>a</b>				
Applications 🕹 Participant Ex	xplorer — 📧 [Termir	al]						10	0:0

🥹 Participant Explorer — Mozilla Firefox	- 0 ×
③ Restore Session × C pgen_int_work/BR5/6× Ø pcdh19_report - Jupg× A cancer	analysis: /m:× 🔯 sact: /main-program × 🔯 rare_diseases_partic × 🕥 Participant Explorer × +
$\leftarrow \rightarrow$ C O A https://prod.terminology-service.aws.gel.ac	☆ 🛛 🗢
Available Columns	Selected Columns
Participant 🛃	Participant 🖻
Ethnic Category Family ID Family Group Type Consent Form	Participant ID     Programme     Proband/Relative     Year of Birth     Date of Death       Phenotypic Sex
Cancer Disease Type       Cancer Disease Sub Type       Rare Disease Group         Rare Disease Sub Group       Specific Rare Disease         Rare Diseases Family Case Report       Interpretation Request ID         Family Case Solved       Additional Comments         Condition / Observation / Procedure       Interpretation         Clinical Concept       Code       Description         Genome Sequence       Image: Concept Code	<ul> <li>Notes:</li> <li>The result contains unique rows only.</li> <li>The result may contain multiple rows per participant, when columns with multiple values are selected. For example, multiple values for Genome Build are rendered as separate rows.</li> <li>Rare Diseases Family Case Report, Condition/Observation/Procedure and Genome Sequence columns cannot be combined in the same result.</li> <li>When Participant ID is not selected, a column is included with the count of participants in the search result for each unique row.</li> </ul>
Genome Build     Platekey     Type     Sample Date     Delivery Date     Delivery ID       Delivery Version     Path	
🚍 Applications 💩 Participant Explorer — 📨 [Terminal]	10:0

Participant Explorer — Mozilla Firefox	- 0 ×
① Restore Session × ◯ pgen_int_work/BRS/S× 🧧 pcdh19_report - Jup∑× 🚺 cancer_ana	alysis: /m:× 🔕 sact: /main-program× 🔕 rare_diseases_partic× 🕐 Participant Explorer × +
$\leftarrow \rightarrow$ C (O A https://prod.terminology-service.aws.gel.ac	☆ ♡ ≡
Available Columns	Selected Columns
Participant       Image: Consent Form         Programme       Proband/Relative       Date of Death       Ethnic Category       Family ID         Family Group Type       Consent Form         Recruited Disease       Image: Cancer Disease Sub Type       Rare Disease Group       Rare Disease Sub Group         Specific Rare Disease       Image: Case Report       Image: Case Report       Image: Case Report	Participant  Participant ID Year of Birth Phenotypic Sex Recruited Disease  Opening participant-search-result-2022-04 X You have chosen to open: Participant-search-result-2022-04-20T09 09 52.csv Which is: CSV document (4.3 KB) from: blob:
Interpretation Request ID       Family Case Solved       Additional Comments         Condition / Observation / Procedure       Image: Code System       Code         Clinical Concept       Code System       Code       Description         Genome Sequence       Image: Code System       Code       Delivery Date       Delivery ID         Delivery Version       Image: Code System       Image: Code System       Code       Delivery ID	What should Firefox do with this file? <ul> <li> Open with </li> <li> GVim (default) </li> <li> iple values are selected. For </li> <li> imple values are selected. For  </li> </ul>

#### Participant Explorer pros/cons

Easy to use point and click interface

Natural language search, don't need to know the codes

Can combine clinical concept filters with participant details such as sex, life status and ethnicity

Many concepts that are disparate in Labkey are consolidated in PXA (eg death, diagnoses)

X Not all of the data in the LabKey tables available to be searched

X Not possible to save searches and return to them for new releases or small tweaks

X Underlying data not exposed in results, making it difficult to verify in bulk

### 4. Labkey tables for cohort building



- Participant details and family relationships
- Sample details
- Genomic file locations
- Clinical data
  - Primary data: GEL
  - Secondary data: NHS, PHE and ONS
- Bioinformatics analysis results



#### Cancer disease

- Recruited disease
  - cancer\_analysis.disease\_type
- Diagnosis codes
  - Hospital episode statistics: hes\_###.diag##
    - Accident and emergency: hes\_ae
    - Admitted patient care: hes\_apc
    - Critical care : hes\_cc
    - Outpatient: hes\_op

Common to use cancer\_analysis.disease\_type to find participants, then verify with diagnosis codes from secondary data

#### Cancer staging

- cancer\_staging\_consolidated
  - TNM (Tumour, Node, Metastasis)
  - AJCC (American Joint Committee on Cancer)
  - Dukes (bowel)
  - Gleason (prostate)
  - FIGO (uterine)
  - HER, ER and PR status (breast)
- cancer\_participant\_tumour, sact and av\_tumour
  - cancer\_staging\_consolidated comprises data from these tables
  - cancer\_staging\_consolidated does not have all participants and you may find more participants that fit your criteria by expanding your search to these tables

#### Cancer treatment

- Systemic anti-cancer therapy: sact
  - analysis\_group all the drugs the participant was treated with
  - drug\_group the drug being referred to in this line of the table
  - Details of how/when the drug was administered

#### Rare disease phenotype

- rare\_disease\_analysis.normalised\_specific\_disease
- rare\_diseases\_participant\_disease
  - Name and categorisation of the disease
- rare\_diseases\_participant\_phenotype
  - HPO term and definition
  - HPO term present?
  - Onset, progression and severity

#### Rare disease cases

If you're looking for rare disease cases that have not been solved yet, or you want to learn more about solved cases (e.g., for eligibility to participate in a clinical trial), you can filter by:

- gmc\_exit\_questionnaire.case\_solved\_family
- death\_details

#### Common disease

- Diagnosis codes
  - Hospital episode statistics: hes\_###.diag##
    - Accident and emergency: hes\_ae
    - Admitted patient care: hes\_apc
    - Critical care : hes\_cc
    - Outpatient: hes\_op

## 5. Covariates in cohort building

#### Covariates to consider

- Age
- Sex
- Ethnicity
- Alive?



- We don't have Age stored
  - You need to calculate it from year\_of\_birth (participant table)
  - Age will always be an approximation, since we only have year and not full date: this is particularly important for anything in small children
- What age do you want to know?
  - Age now (current year participant.year\_of\_birth)
  - Age at sampling (clinic\_sample.clinic\_sample\_datetime participant.year\_of\_birth)
  - Age at diagnosis (hes\_###.###date participant.year\_of\_birth)
  - Age at death (death\_details.death\_date participant.year\_of\_birth



- participant.participant\_phenotypic\_sex: male/female/indeterminate
- participant.participant\_karyotypic\_sex: XX/XY + aneuploidies
- participant.participant\_stated\_gender: male/female/not stated

#### Ethnicity

- participant.participant\_ethnic\_category: what they have ticked on a form
- aggregate\_gvcf\_sample\_stats.pred\_ethnicity\_ancestry: 0-1
  - Ethnicities are 1000 Genomes super-populations: African, South Asian, East Asian, European, American
  - If score  $\geq$  0.8, participant is this population
  - If all scores <0.8, participant is admixed

#### Alive?

- death\_details
- mortality
- cen (cohort event notification)
- av\_patient
- ons (Office of National Statistics)

We can be certain that a participant is dead, but not that they are alive. They could have died since the last data freeze.

# 6. Using the Labkey API in Python and R to build cohorts

#### LabKey API

Labkey API allows you to:

- Combine data and filters from multiple tables
- Work in a variety of programming languages, but most support for Python and R
- Work both locally and on the HPC



#### Set up .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

#### Materials

- Slides and video will be sent out to you after the session
- Scripts available in /gel\_data\_resources/example \_scripts/workshop\_scripts/coh ort\_building\_20220524

cohort_building_2	0220524 - File Manager	-	Х
<u>File Edit View Go He</u>	p		
DEVICES	< cohort_building_20220524 cohort_building_training		>
2.2 MB Volume			 
5.5 MB Volume			
445 MB Volume	cohort building t cohort building t		
485 MB Volume	raining raining.ipynb		
629 MB Volume			
6.0 MB Volume			
PLACES			
n eperry			
Desktop			
🗿 Rubbish Bin			
🧊 genomes			
🧊 gel_data_resou			
🧊 public_data_res			
🧊 re_gecip			
🧊 discovery_forum			
Restricted_access			
📺 pgen_int_work			
	/ ITOMS ( / / / K) Free share / / PK		

#### Accessing the notebooks

Python

#### R

module load python/3.8.1
jupyter notebook --port
<four digit port number>

module load R/4.0.2
rstudio

Open link in browser

## Demo: Labkey API

https://research-help.genomicsengland.co.uk/display/GERE/Participant+Explorer

# 7. Fetching data for downstream analysis

#### Data to include

- genome\_file\_paths\_and\_types
  - gvcfs
  - SV/CNV files
- For filtering out data from gene-centric pipelines (Gene-Variant workflow, gene-centric SNV report)
  - participant\_id
- Phenotype files (for AVT and GWAS downstream analysis):
  - Platekey
  - Phenotype: 0, 1 or score
  - Covariates: age, sex etc

## 8. Getting help and questions

#### Key takeaways

Use Participant Explorer for pointand-click Labkey tables contain loads of data for creating cohorts Copy-and-paste whatever code snippets you need from the notebooks!

#### Getting help



Check our documentation:

- https://research-help.genomicsengland.co.uk/
- Click on the documentation icon in the environment



Contact our Service Desk:

ge-servicedesk@genomicsengland.co.uk

#### Questions



Your microphones are all 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

#### Future sessions

Finding participa genomic data	nts based on		Using the HPC	to run jobs		
	20	Sep	Э.			
22 July			22 Nov.			
Getting medica participants		al his	story for			

## Feedback



## Thank you

