Version: v1
Date: March 22, 2022
Author: Anna Vavříková (avavrikova@dnanexus.com)

HDRUK UKB 500K Pheno/Geno(array) 1.0

Powered by DNAnexus

User Acceptance Testing
March 22, 2022

Document Prepared By:
xVantage Group

Version: v1
Date: March 22, 2022

Introduction and Setup

This document contains a series of test scripts intended for the user acceptance testing (UAT) of the DNAnexus Apollo environment on HDRUK UKB 500K pheno/geno(array) 1.0 data.

HDRUK has nominated a list of experts to evaluate the system; they will be referred to as Evaluators throughout this document. The evaluators are:

- Adam Butterworth (asb38@medschl.cam.ac.uk)
- Samuel Lambert (sl925@medschl.cam.ac.uk)
- Xiyun Jiang (xj259@cam.ac.uk)
- David Stacey (ds763@medschl.cam.ac.uk)
- Martin Kelemen (mk907@medschl.cam.ac.uk)
- Stephen Kaptoge (skk22@medschl.cam.ac.uk)
- Yi Yang (yyy30@medschl.cam.ac.uk)
- Jing Hua Zhao (jhz22@medschl.cam.ac.uk)

In order to perform the test scenarios in this document, several setup steps must be performed on the DNAnexus platform.

Prerequisites

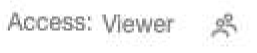
DNAnexus has created and set up a UAT project for you. As a result, you should:

- Have access to DNAnexus. If you do not have access, please reach out to your internal project manager to have an account created.
- Be a member of the HDRUK designated organization:
  - **org-hdruk_multiomics**.
- Have Viewer (or higher) access to the project created for the Acreated for the UKB 500k Pheno/Geno(array) project:
  - **UK Biobank - HDRUK London UAT**
- Open the dataset for UAT within the Cohort Browser:
  - **hdruk_500k_ukb_merged_pheno_geno**

Check access to Apollo Early Access Datasets

To confirm that you have access, please:

1. Log into the DNAnexus platform
2. In the projects list, find the project called **UK Biobank - HDRUK London UAT**.
3. In the upper right corner of the page, look at the access level. It should say **Access: Viewer** as in this screenshot:



Access: Viewer

If you do not have access, please let the UAT lead know to fix the issue.

Confirm Receipt of Apollo Early Access Datasets

□ Access to the UKB 500K Pheno/Geno(array) dataset.

Signature 1

I, a HDRUK Apollo Early Access Evaluator, confirm access to the above data.

□ Accept	□ Conditionally Accept (conditional criteria below)	□ Reject
Signature:		Date:
Name (written):		

Conditional Acceptance:

Test Scenarios - Cohort Browser

In the following scenarios, we will perform various operations using the Apollo Cohort Browser - an app with dashboard-like graphical user interface (GUI or UI) that allows exploring Apollo datasets and underlying databases, filtering, searching, and visualizing of data, creating dashboard and Cohort Objects, and collaborating by sharing these objects.

For more information please see our help documentation:
https://documentation.dnanexus.com/user/cohort-browser

Scenario 1. Dataset exploration using Cohort Browser

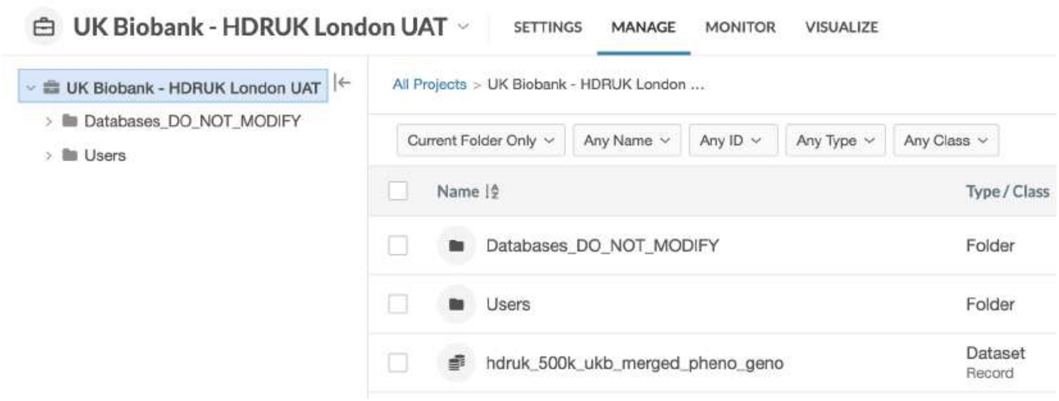
Objective

In this scenario, we will explore a dataset using the Cohort Browser. We will add and remove data tiles to the dashboard, and save it as a new dashboard.

Start Cohort Browser

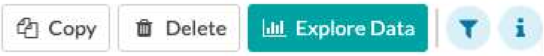
The Cohort Browser is a web application built on the DNAnexus platform. It runs within a web browser. To start the cohort browser please:

1. Login to the DNAnexus platform and open the **UK Biobank - HDRUK London UAT** project.
2. Select the **hdruk_500k_ukb_merged_pheno_geno** object in the root folder.



Name	Type / Class
Databases_DO_NOT_MODIFY	Folder
Users	Folder
hdruk_500k_ukb_merged_pheno_geno	Dataset Record

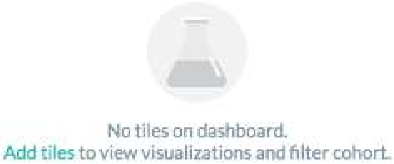
3. Click the **Explore Data** button in the toolbar. This will launch a new browser window (or tab) and start the Cohort Browser. (Or you can click on the dashboard name itself.)



Add and remove tiles

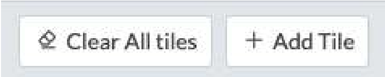
In this exercise, we will add a data field of interest as a tile and remove tiles that are not needed.

1. After completing the steps above, the cohort browser should now be running. If this is your first time opening the cohort browser, the dashboard will most likely be empty.
2. If you see any tiles populated within the dashboard, empty the dashboard by clicking the **Clear All tiles** button on the right side or clicking the **X** in the top right corner of every tile. You will know when the dashboard is empty when you see this message in the dashboard:



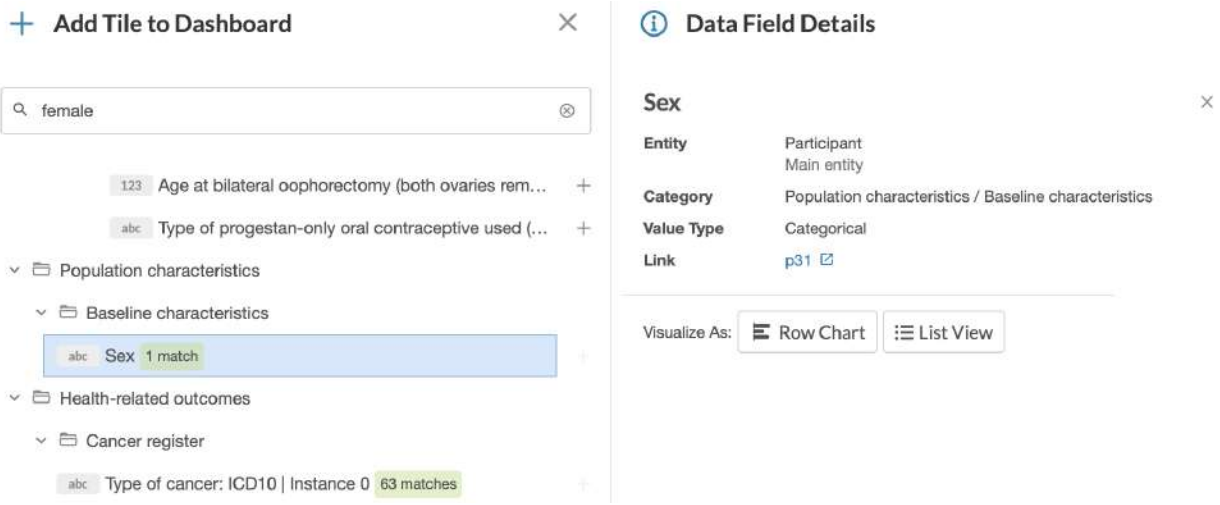
No tiles on dashboard.
Add tiles to view visualizations and filter cohort.

3. To populate the dashboard, Click + **Add tile** button in the upper right corner.



This will reveal a sidebar on the left side containing a hierarchical view of all the fields present within the dataset. The UK Biobank hierarchy of fields have multiple layers. You can browse for a field of interest by unfolding each folder. It is often easier to find a field of interest by typing a keyword into the Search field.

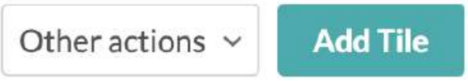
4. To locate the folder that contains the Sex of a UKB participant, type “**female**” into the search box. Scroll down through the search results until you locate the **Sex** at the bottom of the field.


Add Tile to Dashboard
female
123 Age at bilateral oophorectomy (both ovaries rem...
abc Type of progestan-only oral contraceptive used (...
Population characteristics
Baseline characteristics
abc Sex 1 match
Health-related outcomes
Cancer register
abc Type of cancer: ICD10 | Instance 0 63 matches
Data Field Details
Visualize As:
Row Chart
List View

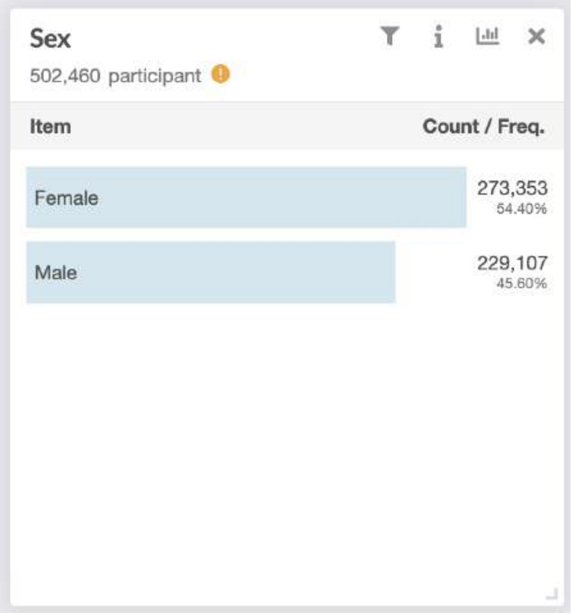
Sex

Entity	Participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31

5. Click **Add as Tile** button to add the field to your dashboard.

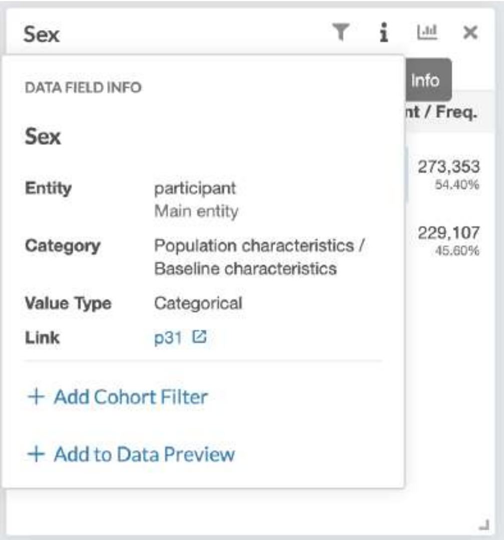


6. Click away from the sidebar by clicking on the **x** in the top right corner. With the side bar closed, you should now be able to see your dashboard with your first tile.


Sex
502,460 participant

Item	Count / Freq.
Female	273,353 54.40%
Male	229,107 45.60%

You can invoke tile Data Field Info pop-up by click the ℹ icon in the upper right corner of each tile. When you look at the **Category** section of the tile, you can see that Baseline Characteristics of the UKB Population are under the **Population Characteristics folder**.



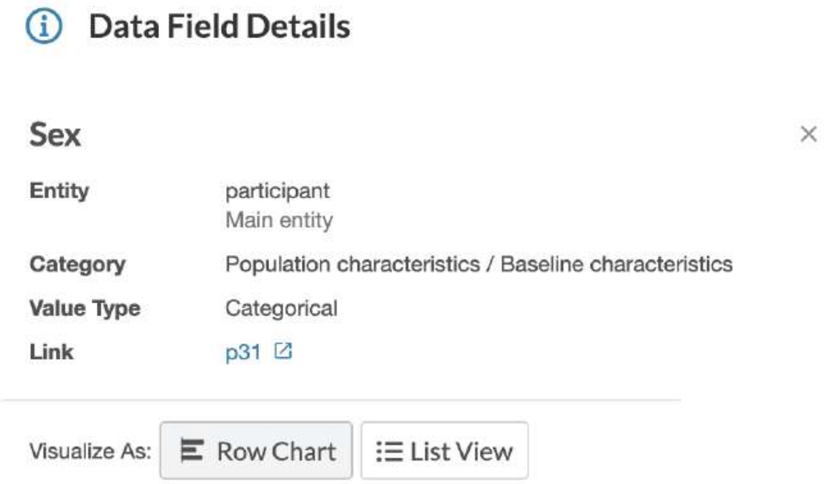
DATA FIELD INFO

Sex

Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31

The link p31 will take you to the UK Biobanks showcase which contains more metadata about each field.

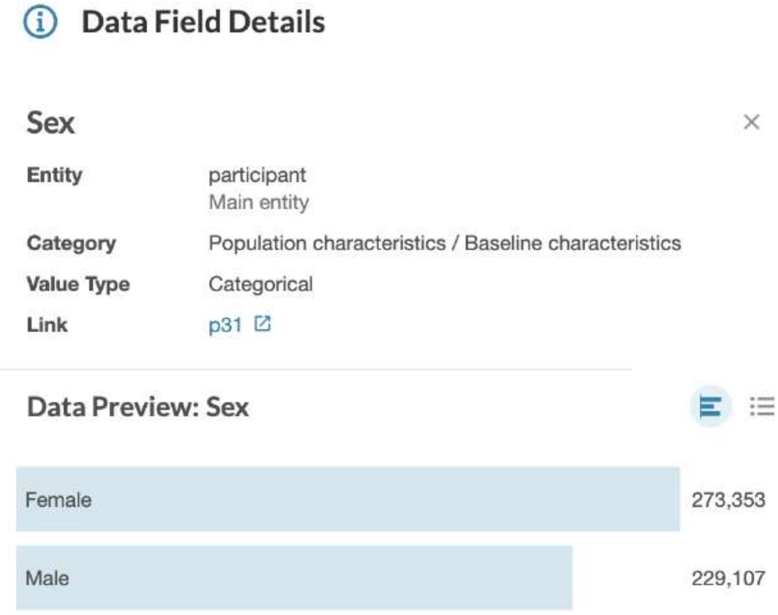
NOTE* - You also have the option to view the field distribution before adding the field as a tile. This can be done by clicking on the **Row chart or **List View** buttons (shown below), after selecting your field of interest from the side bar.



Data Field Details

Sex

Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31



Data Field Details

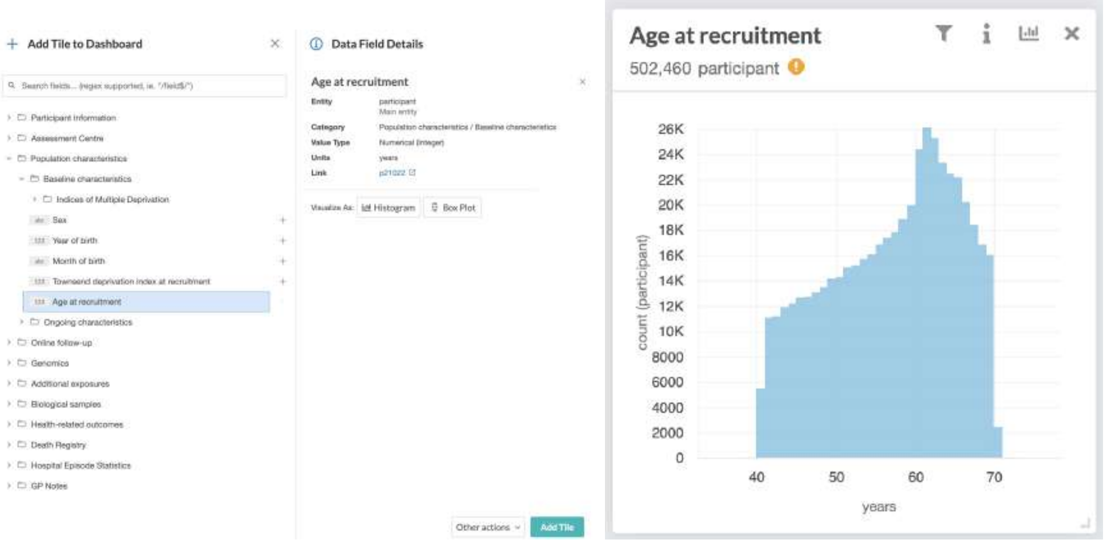
Sex

Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31

Data Preview: Sex

Female	273,353
Male	229,107

7. We can now look at **Age at recruitment,** to do this navigate to the sidebar again, by clicking on **Add Tile.** Unfold the **Population characteristics** folder and unfold **Baseline characteristics** folder and add the **Age at recruitment** tile.



Data Field Details

Age at recruitment

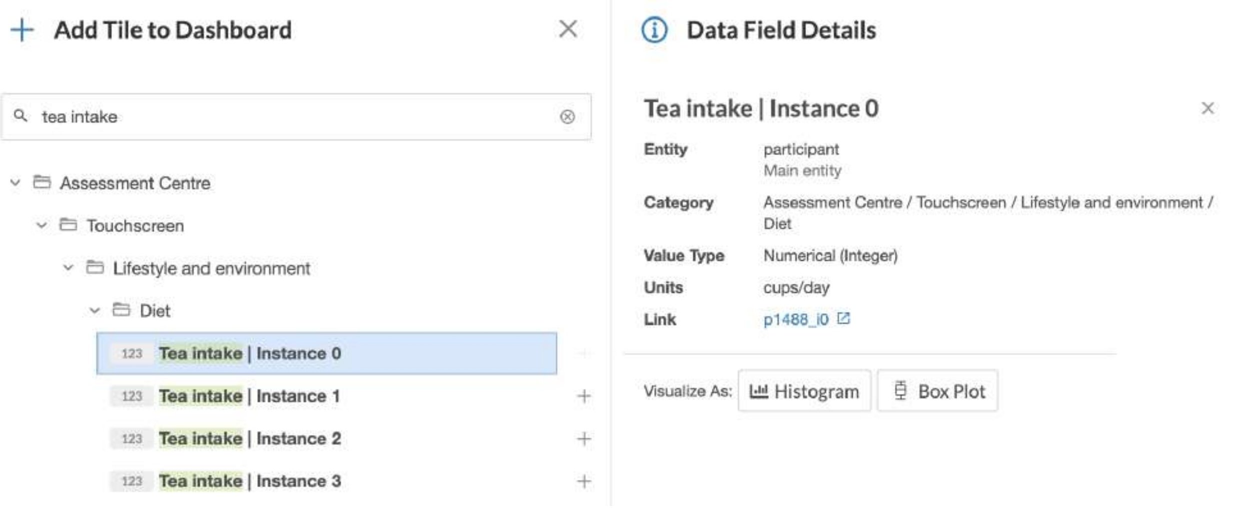
Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Numerical (integer)
Units	years
Link	p21022

Age at recruitment

502,460 participant

Hover over the ❗ icon to see the completeness of the field. You can see that there is one missing value for both **Sex** and **Age at recruitment**.

8. Navigate to the **Tea intake tile** (in folder Assessment Centre > Touchscreen > Lifestyle and environment > Diet) and add **Tea intake | Instance 0** to the dashboard:


Add Tile to Dashboard
tea intake
Assessment Centre
Touchscreen
Lifestyle and environment
Diet
123 Tea intake | Instance 0
123 Tea intake | Instance 1
123 Tea intake | Instance 2
123 Tea intake | Instance 3
Data Field Details
Visualize As:
Histogram
Box Plot

Tea intake | Instance 0

Entity participant
Main entity

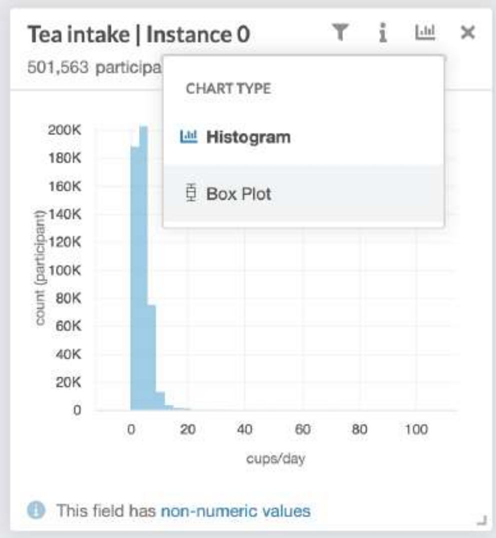
Category Assessment Centre / Touchscreen / Lifestyle and environment / Diet

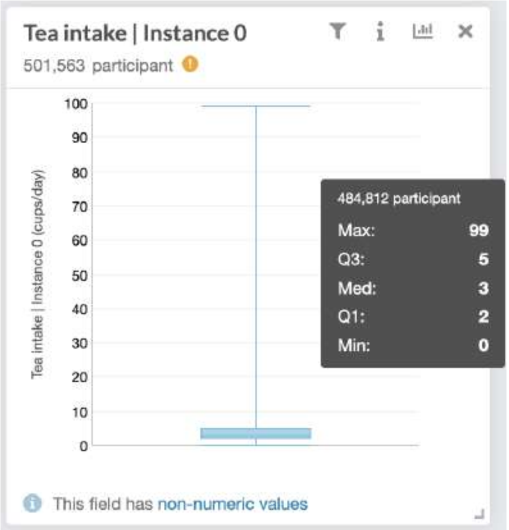
Value Type Numerical (Integer)

Units cups/day

Link p1488_i0

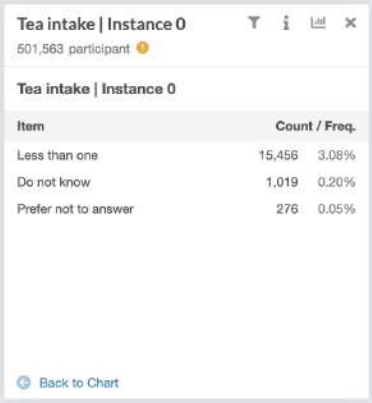
9. Adjust the chart type for **Tea intake | Instance 0** by clicking on the icon shaped like a vertical bar chart.





By hovering over the boxplot you can get basic statistics for this field.

ⓘ This field has non-numeric values message tells you that there are some coded values for this field. By clicking the link you will get a list view of the coded values.



Tea intake | Instance 0
501,563 participant

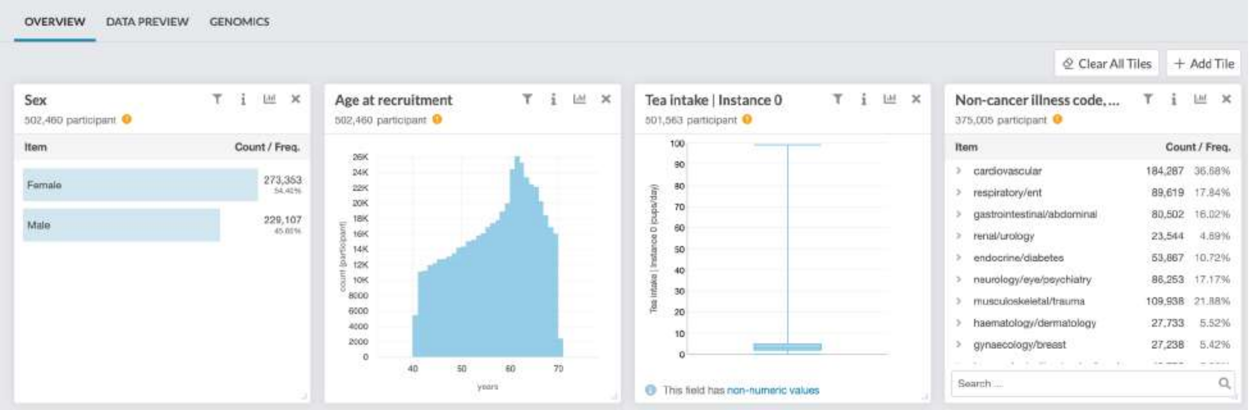
Tea intake | Instance 0

Item	Count / Freq.	
Less than one	15,456	3.08%
Do not know	1,019	0.20%
Prefer not to answer	276	0.05%

Back to Chart

10. For practice, add following tiles:
    a. **Date of death** (Death registry > Death)
    b. **Non-cancer illness code, self-reported | Instance 0** (Assessment Centre > Verbal interview > Medical conditions)

11. Remove the **Date of death** tile by clicking the **X** button in the upper right corner of the tile.
    Your browser should appear similar to the image below:



Sex
502,460 participant

Item	Count / Freq.
Female	273,353 54.40%
Male	229,107 45.60%

Age at recruitment
502,460 participant

Tea intake | Instance 0
501,563 participant

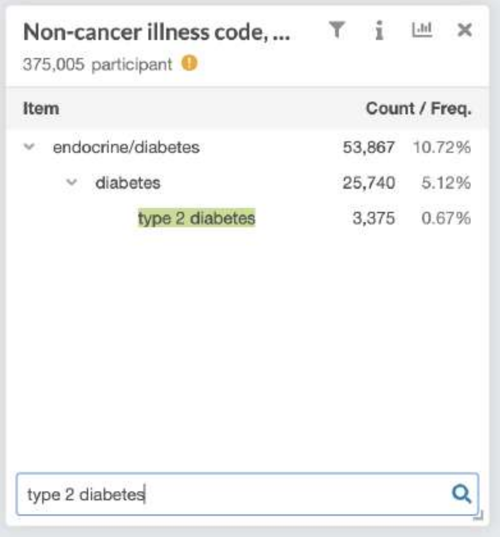
This field has non-numeric values

Non-cancer illness code, ...
375,005 participant

Item	Count / Freq.	
cardiovascular	184,287	36.68%
respiratory/ent	89,619	17.84%
gastrointestinal/abdominal	80,502	16.02%
renal/urology	23,544	4.69%
endocrine/diabetes	53,867	10.72%
neurology/eye/psychiatry	86,253	17.17%
musculoskeletal/trauma	109,938	21.88%
haematology/dermatology	27,733	5.52%
gynaecology/breast	27,238	5.42%

Search ...

Non-cancer illness code, self-reported | Instance 0 is a hierarchical categorical field. You can search for your value of your interest with the search feature at the bottom of the tile. Try searching for **type 2 diabetes**.



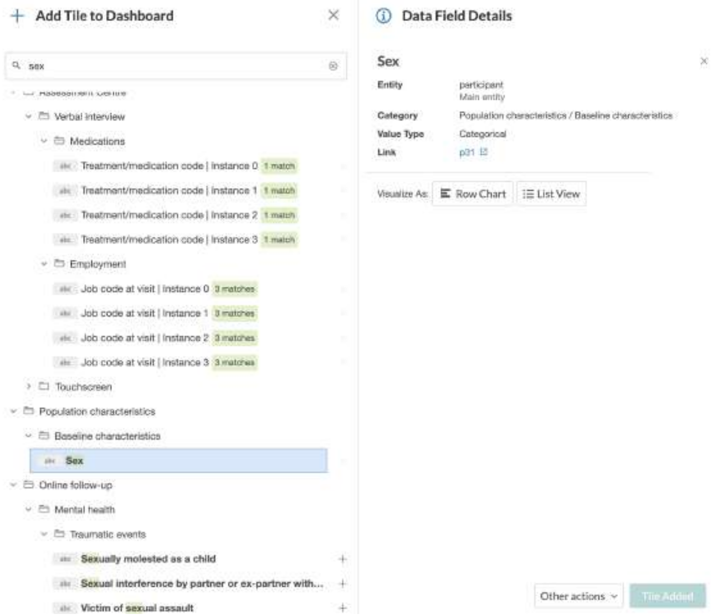
Non-cancer illness code, ...

375,005 participant

Item	Count	Freq.
endocrine/diabetes	53,867	10.72%
diabetes	25,740	5.12%
type 2 diabetes	3,375	0.67%

type 2 diabetes

12. If desired, drag the bottom right corner of each tile to adjust the size of each tile. You can reshuffle tiles by dragging and dropping them to different locations.
13. You can also create 2 dimensional tiles. We will create **Sex x Body mass index (BMI) | Instance 0** tile. Search for field **Sex**, one you find it, click on it.



Add Tile to Dashboard

sex

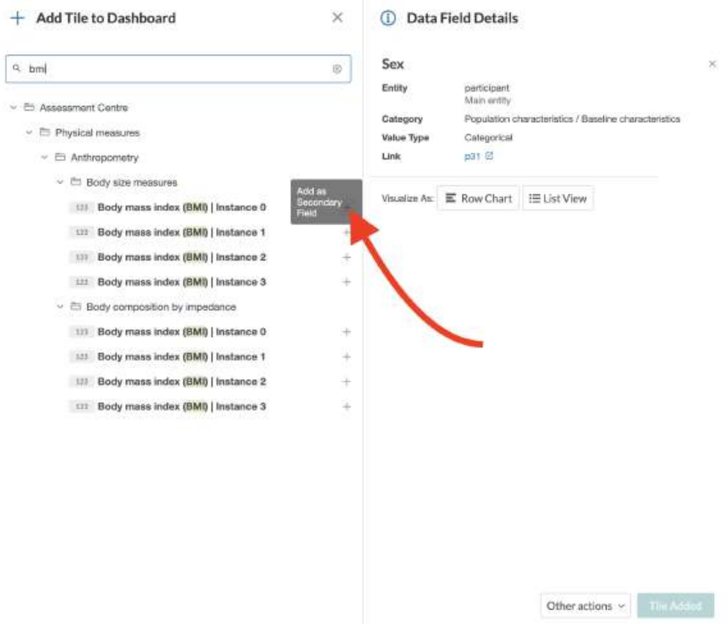
- Assessment Centre
  - Verbal interview
    - Medications
      - Treatment/medication code | Instance 0 1 match
      - Treatment/medication code | Instance 1 1 match
      - Treatment/medication code | Instance 2 1 match
      - Treatment/medication code | Instance 3 1 match
    - Employment
      - Job code at visit | Instance 0 3 matches
      - Job code at visit | Instance 1 3 matches
      - Job code at visit | Instance 2 3 matches
      - Job code at visit | Instance 3 3 matches
  - Touchscreen
- Population characteristics
  - Baseline characteristics
    - Sex
- Online follow-up
  - Mental health
    - Traumatic events
      - Sexually molested as a child
      - Sexual interference by partner or ex-partner with...
      - Victim of sexual assault

Data Field Details

Sex

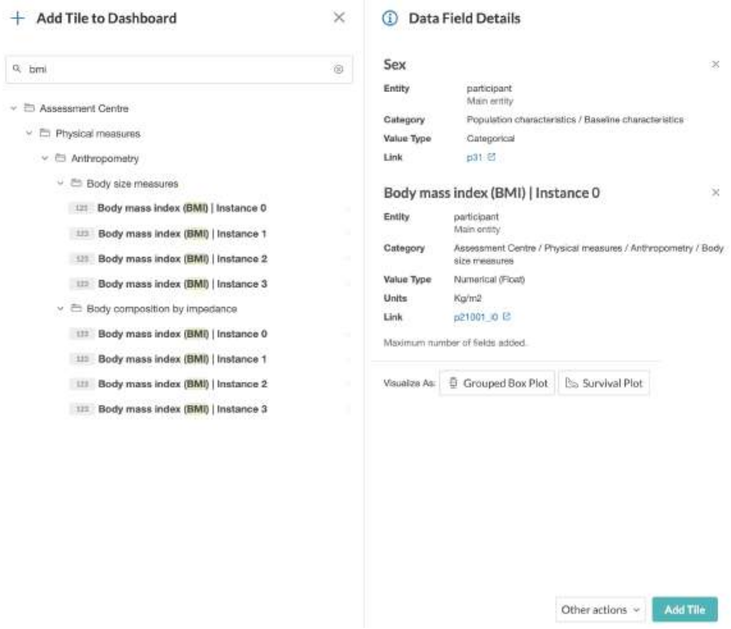
Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31

Start searching for the **Body mass index (BMI) | Instance 0.** Without clicking anywhere else, click on the plus sign next to the **Body mass index (BMI) | Instance 0** field. You will see that the Data Field Details section changed. Click **Add Tile** button. You can find more details about how to build multidimensional plots in our documentation.


Add Tile to Dashboard
bmi
Assessment Centre
Physical measures
Anthropometry
Body size measures
Body mass index (BMI) | Instance 0
Add as Secondary Field
Body mass index (BMI) | Instance 1
Body mass index (BMI) | Instance 2
Body mass index (BMI) | Instance 3
Body composition by impedance
Body mass index (BMI) | Instance 0
Body mass index (BMI) | Instance 1
Body mass index (BMI) | Instance 2
Body mass index (BMI) | Instance 3
Data Field Details
Visualize As:
Row Chart
List View
Other actions
Tile Added

Sex

Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31


Add Tile to Dashboard
bmi
Assessment Centre
Physical measures
Anthropometry
Body size measures
Body mass index (BMI) | Instance 0
Body mass index (BMI) | Instance 1
Body mass index (BMI) | Instance 2
Body mass index (BMI) | Instance 3
Body composition by impedance
Body mass index (BMI) | Instance 0
Body mass index (BMI) | Instance 1
Body mass index (BMI) | Instance 2
Body mass index (BMI) | Instance 3
Data Field Details
Visualize As:
Grouped Box Plot
Survival Plot
Other actions
Add Tile

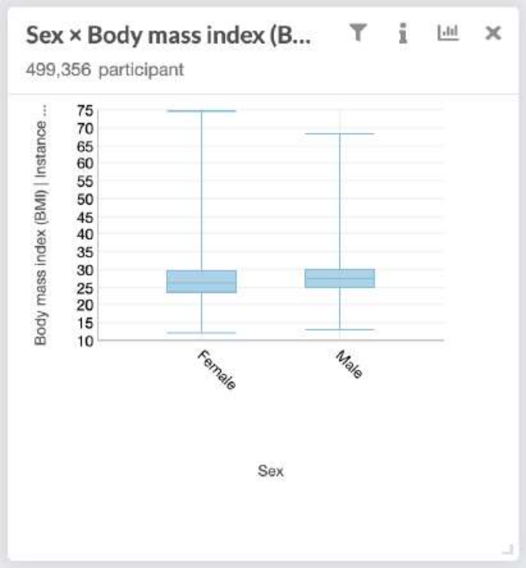
Sex

Entity	participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31

Body mass index (BMI) | Instance 0

Entity	participant Main entity
Category	Assessment Centre / Physical measures / Anthropometry / Body size measures
Value Type	Numerical (Float)
Units	Kg/m2
Link	p21001_i0

Maximum number of fields added.

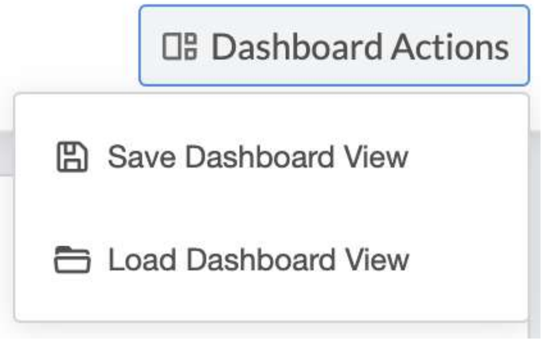


**Note The maximum number of tiles on the dashboard is 15. Once you reach it, you will not be able to add more tiles.*

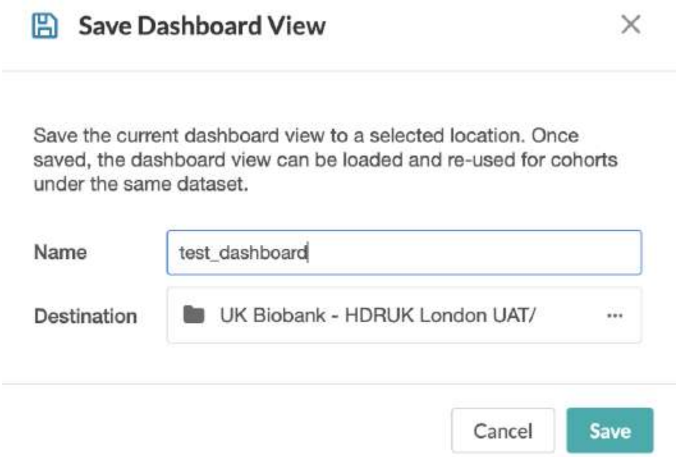
Save dashboard view

It is possible to save your favorite dashboard view, to use later, or to share.To save your dashboard view:

1. Open the **Dashboard Actions** dropdown menu, in the upper right corner, and select **Save Dashboard View**:



2. In the dialog box, enter **test_dashboard** as the Name. The dialog box should now look like this:



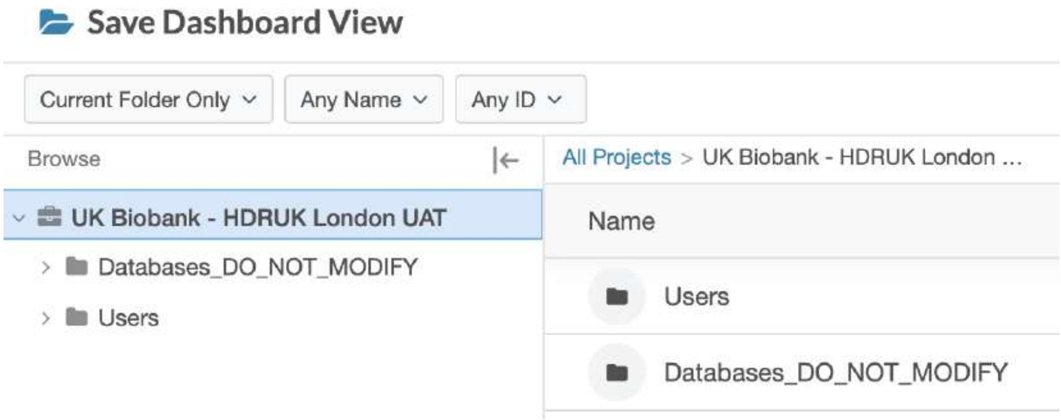
Save the current dashboard view to a selected location. Once saved, the dashboard view can be loaded and re-used for cohorts under the same dataset.

Name: test_dashboard

Destination: UK Biobank - HDRUK London UAT/

3. Click the browse field next to the Destination. Browse to your test project's root folder, navigate to your user folder and click the **Select Folder** button.

   NOTE* - If you would like to save your dashboard in a different project, you may need to navigate to **All Projects in order to view a list of all projects you have access to.



4. Click **Save**. The dashboard object will be saved in your user folder.

Open previously saved dashboard view

1. Clear all tiles on your current dashboard.
2. In the open Cohort Browser window that was used in the tasks above (i.e. that loaded the UKB 500k Pheno/Geno(array) dataset), open the **Dashboard Actions** dropdown menu and select **Load Dashboard View.**

3. Browse to your user folder within the project, select the previously saved **test_dashboard** and click **Select Data**. This will load the tiles from your saved dashboard view.

Feature checklist

I confirm that I was able to:

- ▢ Start Cohort Browser for the UKB 500k Pheno/Geno(array) dataset
- ▢ Add and remove tiles from my dashboard
- ▢ Save my dashboard
- ▢ View saved dashboard

Signature 2

I, a HDRUK Apollo Early Access Evaluator, have performed the actions described above and confirm the features were available and tested by myself. For the scenario and the features described I:

▢ Accept	▢ Conditionally Accept (conditional criteria below)	▢ Reject
Signature:		Date:
Name (written):		

Conditional Acceptance:

Scenario 2. Creating and using cohorts

Objective

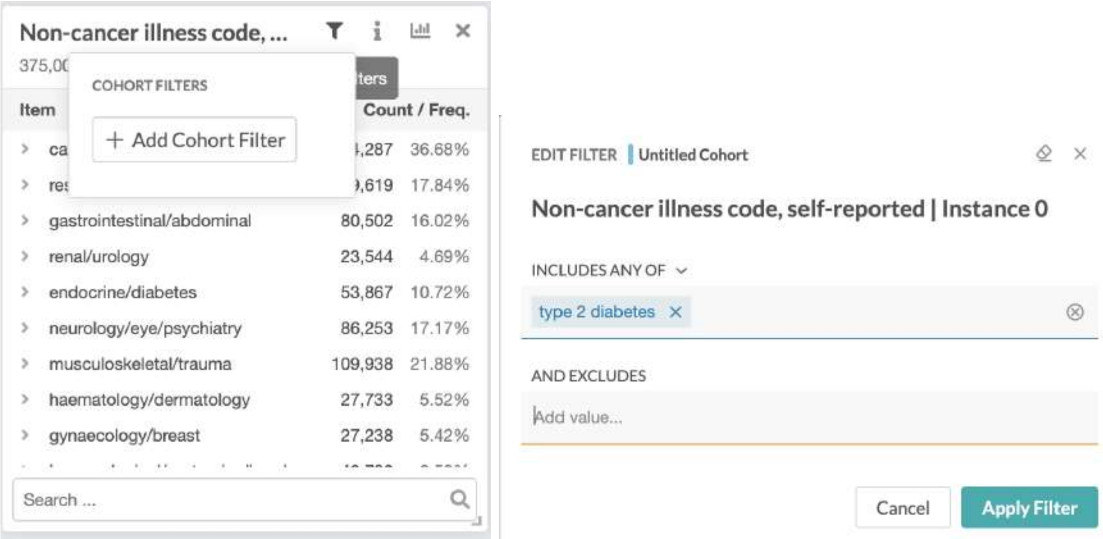
In this scenario, you will explore the hdruk_ukb_500k_pheno_geno dataset to answer the question: **How many Participants in the UKB 500k Pheno/Geno(array) Cohort self-reported themselves as diabetic (type 2 diabetes) and have BMI over 30 kg/m^2? And for that patient population, what is their distribution of types of Treatments (Medication Codes), VLDL cholesterol and phenylalanine?** You will then save these data as a Cohort Object that can be revisited later. Finally, you will share the cohort with a colleague in the UAT test team.

This demonstrates how to filter the dataset using tiles, how to save selected tiles and dashboard layout as cohorts, and how to share them.

Filter the dataset

1. Open **hdruk_500k_ukb_merged_pheno_geno** dataset (Refer to chapter **Start Cohort Browser** above.) Empty the dashboard view by clicking the **Clear All tiles** button if populated.
2. Add tiles to your dashboard, the best way to find these fields is by typing each field name into the search box:
    a. **Non-cancer illness code, self-reported | Instance 0**
    b. **Diabetes diagnosed by doctor | Instance 0**
    c. **Treatment/medication code | Instance 0**
    d. **VLDL Cholesterol | Instance 0**
    e. **Phenylalanine | Instance 0**
    f. **Body mass index (BMI) | Instance 0**

3. Click the filter icon on the **Non-cancer illness code, self-reported | Instance 0** tile then click **Add Cohort Filter**. In the **Edit Filter** menu start typing **type 2 diabetes**. The autocomplete utility should help you with finding the desired field. After that click **Apply Filter**. You should see the Cohort browser updating.



Non-cancer illness code, ...

375,00

Item	Count / Freq.	
ca	,287	36.68%
res	,619	17.84%
gastrointestinal/abdominal	80,502	16.02%
renal/urology	23,544	4.69%
endocrine/diabetes	53,867	10.72%
neurology/eye/psychiatry	86,253	17.17%
musculoskeletal/trauma	109,938	21.88%
haematology/dermatology	27,733	5.52%
gynaecology/breast	27,238	5.42%

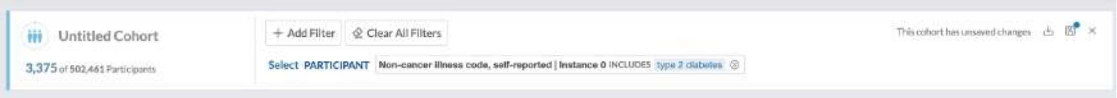
EDIT FILTER | Untitled Cohort

Non-cancer illness code, self-reported | Instance 0

INCLUDES ANY OF: type 2 diabetes

AND EXCLUDES: Add value...

Look in the cohort summary section towards the top of the page to inspect your query.



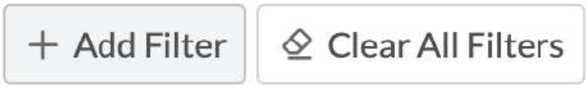
Untitled Cohort

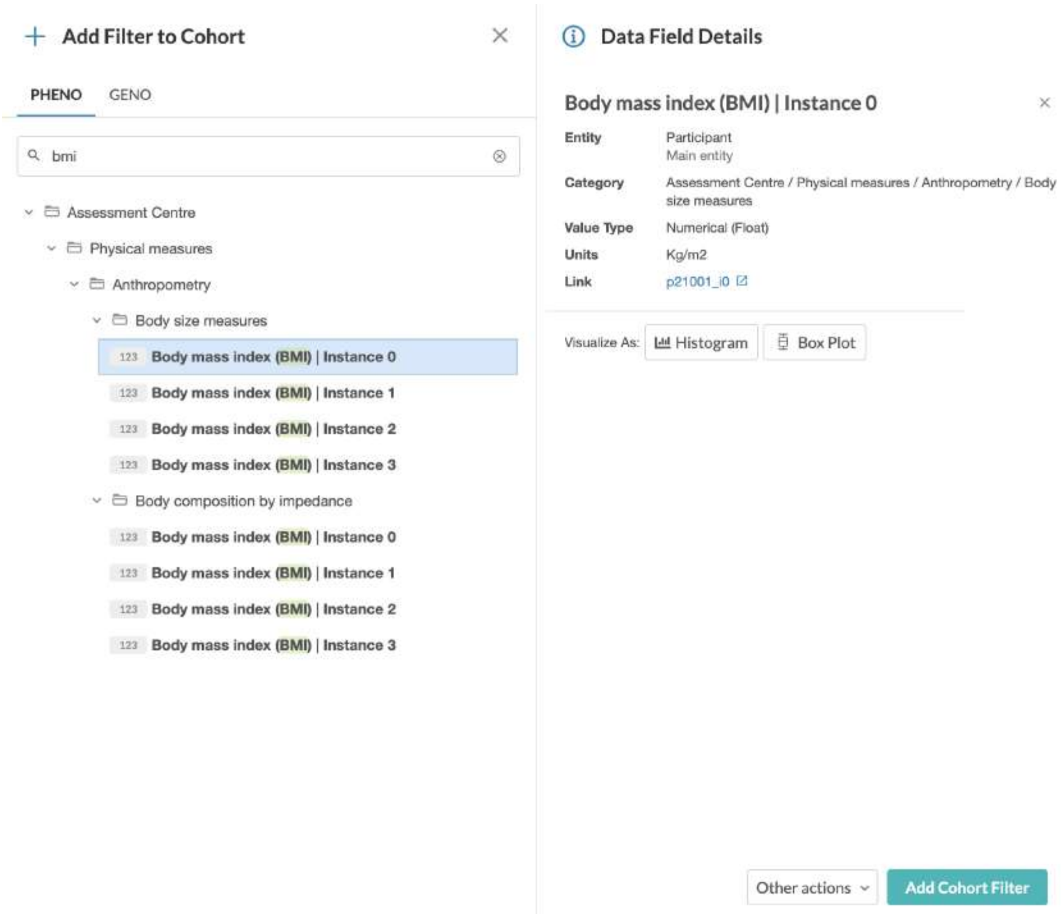
3,375 of 502,461 Participants

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes

This cohort has unsaved changes

4. We will now add a second filter, this time on numerical field: **Body mass index (BMI) | Instance 0**. You can add filters by clicking the **+ Add Filter** button in the cohort summary section. This will guide you to the sidebar menu where you can look for the field. After finding the desired field click the **Add Cohort Filter** button. You should get to the edit filter menu, choose **IS GREATER THAN** filter and set it to **30**.



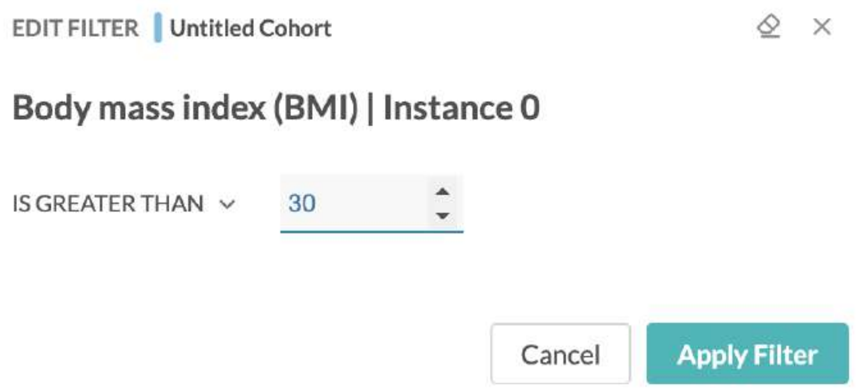


bmi

- Assessment Centre
  - Physical measures
    - Anthropometry
      - Body size measures
        - 123 Body mass index (BMI) | Instance 0
        - 123 Body mass index (BMI) | Instance 1
        - 123 Body mass index (BMI) | Instance 2
        - 123 Body mass index (BMI) | Instance 3
      - Body composition by impedance
        - 123 Body mass index (BMI) | Instance 0
        - 123 Body mass index (BMI) | Instance 1
        - 123 Body mass index (BMI) | Instance 2
        - 123 Body mass index (BMI) | Instance 3

Body mass index (BMI) | Instance 0

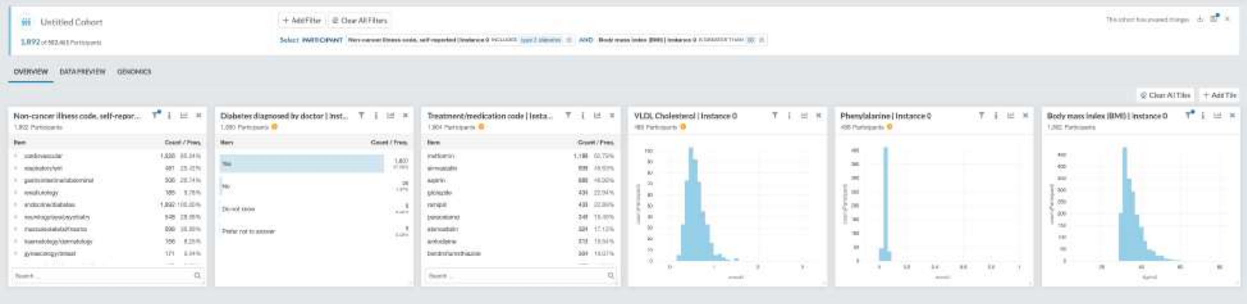
Entity	Participant Main entity
Category	Assessment Centre / Physical measures / Anthropometry / Body size measures
Value Type	Numerical (Float)
Units	Kg/m2
Link	p21001_i0



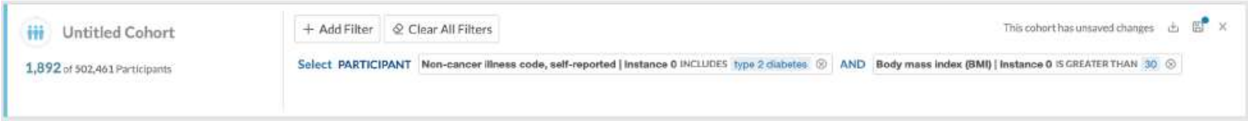
Body mass index (BMI) | Instance 0

IS GREATER THAN 30

Your dashboard should now look like this.

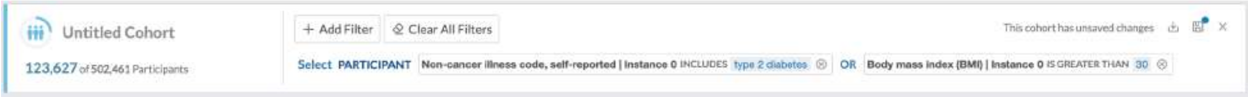

Untitled Cohort
1,892 of 502,461 Participants
+ Add Filter
Clear All Filters
OVERVIEW
DATA PREVIEW
GENOMICS
Clear All Tiles
+ Add Tile
Non-cancer illness code, self-repor...
Diabetes diagnosed by doctor | Inst...
Treatment/medication code | Insta...
VLDL Cholesterol | Instance 0
Phenylalanine | Instance 0
Body mass index (BMI) | Instance 0

5. By default the Cohort Browser creates an AND query between these 2 filters. If you wanted to change this to an OR query you could click on the AND widget to toggle to an OR. However, for the purposes of this scenario make sure you have an AND query. You can learn more about cohort filtering in our documentation.



1,892 of 502,461 Participants

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes AND Body mass index (BMI) | Instance 0 IS GREATER THAN 30



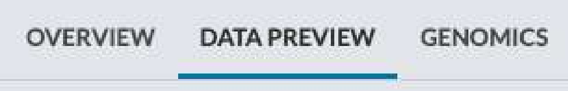
123,627 of 502,461 Participants

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes OR Body mass index (BMI) | Instance 0 IS GREATER THAN 30

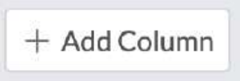
Export data to third party applications

The Cohort Browser table can be used to review individual values of interest and download tables of data as CSV files that can be opened in third party spreadsheet applications.

1. Navigate to the **DATA PREVIEW (**cohort table) section.

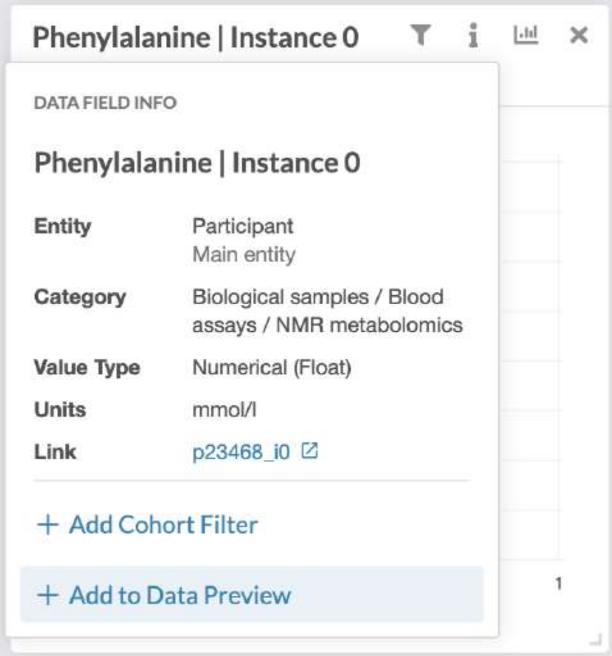


2. Add columns by clicking **+ Add Column**. This will reveal a scrolling sidebar on the left side containing a hierarchical view of all the fields present within our dataset.



NOTE - The **Participant ID** is a default column and cannot be removed.

3. Add the following columns to your table: **Treatment/medication code | Instance 0, VLDL Cholesterol | Instance 0** and **Phenylalanine | Instance 0**. You can add columns by choosing them from the sidebar menu which you can invoke as previously mentioned. You can also add a field to the preview from the field tile, by clicking the **i** icon and selecting **+ Add to Data Preview.**



Phenylalanine | Instance 0

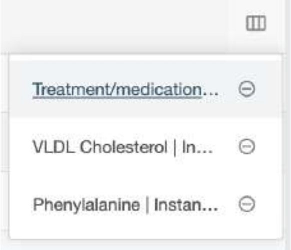
DATA FIELD INFO

Phenylalanine | Instance 0

Entity	Participant Main entity
Category	Biological samples / Blood assays / NMR metabolomics
Value Type	Numerical (Float)
Units	mmol/l
Link	p23468_i0

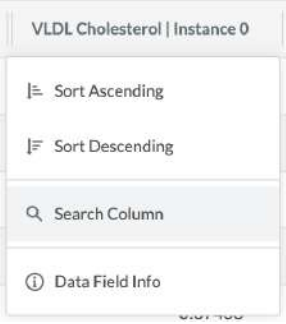
NOTE - the data within the columns may be **Null. That is because some patients within the defined patient population may not have had data collected for the displayed field/column.*

4. To remove a column **Treatment/medication code | Instance 0**, click on the ▥ icon and and select the remove ⊖ icon next to the name of the column.

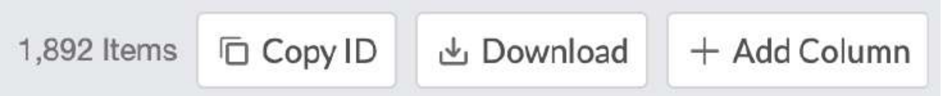


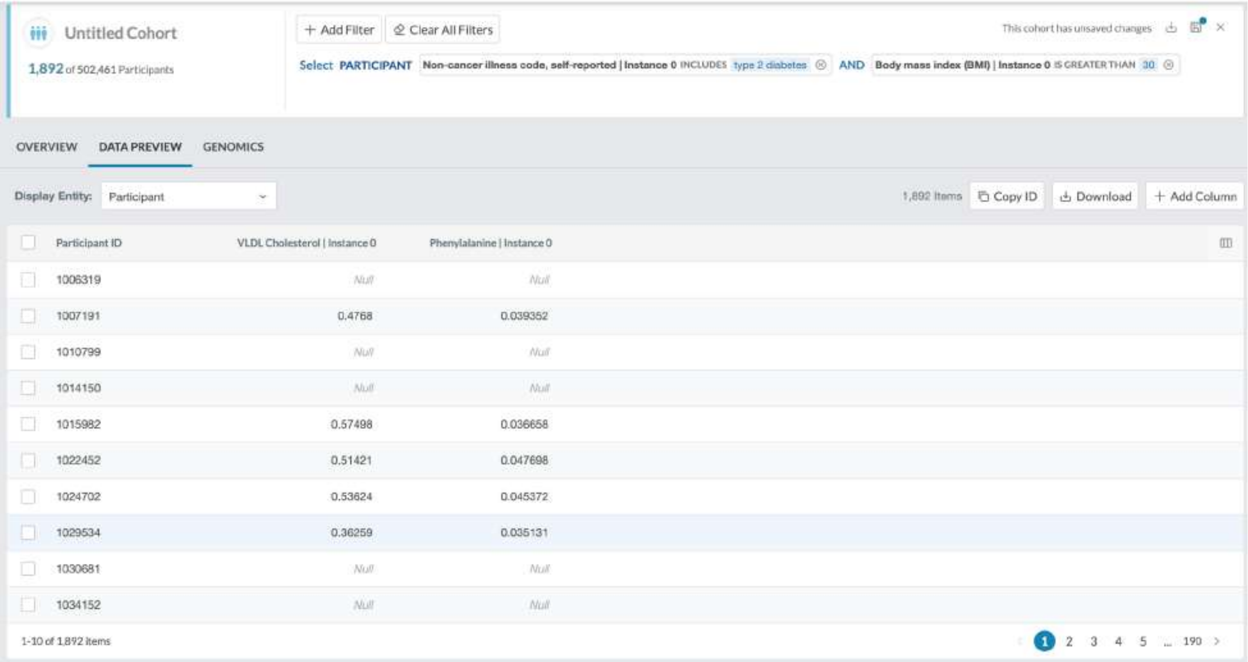
Treatment/medication... ⊖
VLDL Cholesterol | In... ⊖
Phenylalanine | Instan... ⊖

5. Click on the title of each column to access the **sort, search** and **Data Field Info** features.



6. Click on the checkbox on the left of the Participant ID header. All rows will be selected. (Feel free to select just rows of your interest by selecting them individually or by filtering them.)
7. To export the full table in CSV format, click on the **Download** button in the upper right corner of the Data Preview section. The full table will be downloaded to your computer and can be used in third party spreadsheet applications.





1,892 of 502,461 Participants

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes AND Body mass index (BMI) | Instance 0 IS GREATER THAN 30

Display Entity: Participant

□	Participant ID	VLDL Cholesterol | Instance 0	Phenylalanine | Instance 0
□	1006319	*Null*	*Null*
□	1007191	0.4768	0.039352
□	1010799	*Null*	*Null*
□	1014150	*Null*	*Null*
□	1015982	0.57498	0.036658
□	1022452	0.51421	0.047698
□	1024702	0.53624	0.045372
□	1029534	0.36259	0.035131
□	1030681	*Null*	*Null*
□	1034152	*Null*	*Null*

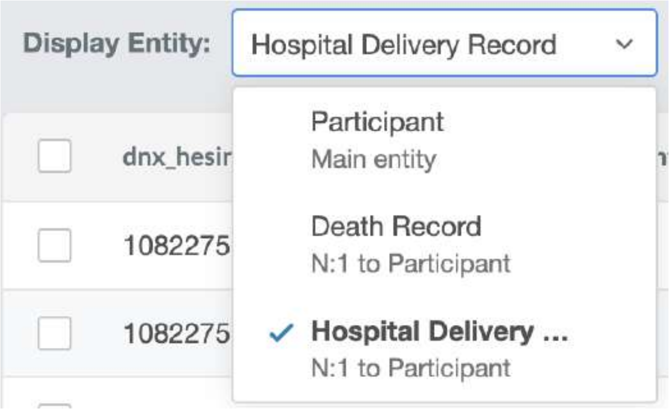
1-10 of 1,892 items

1 2 3 4 5 ... 190

**NOTE The Download button in the DATA PREVIEW tab will export up to 30,000 rows. . To download a table with more than 30,000 rows, use the Table exporter App.*

**NOTE: We strongly recommend not adding more than 30 columns. Please use the [Table exporter App](#) to export larger quantities of data.*

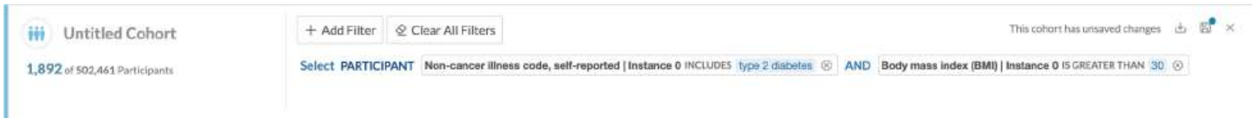
NOTE Your dataset has multiple entities (Patient entity, Death registry entities, Hospital Episode Statistics entities, GP notes entities). In order not to mix columns across entities, you can select columns from relevant entities by toggling between entities using the **Display entity dropdown menu. Each entity table will be downloaded separately, one at a time.*



Save Cohort

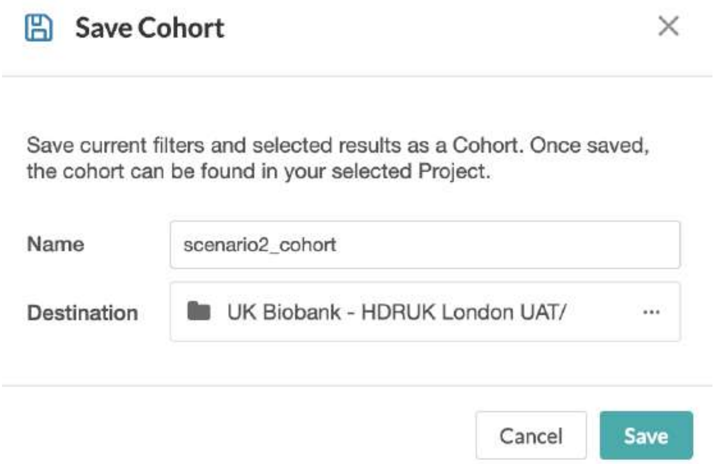
Filtering criteria of a dataset as well as the layout of a dashboard can be stored in a special Cohort object. To do so:

1. Click the disk icon (if you move your over the disk icon it should say **Save cohort**) in the bar where your cohort filters are listed:


Untitled Cohort
1,892 of 502,461 Participants
+ Add Filter
Clear All Filters
This cohort has unsaved changes

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes AND Body mass index (BMI) | Instance 0 IS GREATER THAN 30

2. In the dialog box, enter **scenario2_cohort** as the Name.



Save Cohort

Save current filters and selected results as a Cohort. Once saved, the cohort can be found in your selected Project.

Name: scenario2_cohort

Destination: UK Biobank - HDRUK London UAT/

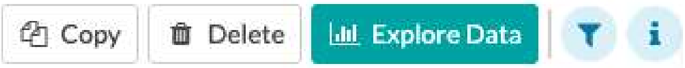
3. Confirm the save location is your user folder.
4. Click **Save**. The dashboard object will be saved in your user folder in the test project.

NOTE you can rename your cohort in the UI by clicking the "Untitled cohort" text field. New name will be suggested next time in the **Save Cohort menu.*

Open Cohort

Cohort objects can be explored similar to Dataset objects. To explore the Cohort Object you created in the previous exercise:

1. Navigate to your user folder in the test project.
2. Select Cohort object called **scenario2_cohort.**
3. Click the **Explore Data** button in the toolbar. You can also do so by clicking the cohort name.



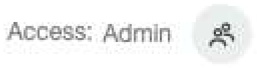
This will launch a new browser window (or tab) and load the Cohort Browser.
You will see that the filtering criteria and the dashboard layout will be exactly the same as it was when you saved the cohort.

Share Cohort

In the DNAnexus platform, the only shareable object is a Project; one cannot share a single file or any other object including a cohort. Instead, one has to share a whole Project containing the object.

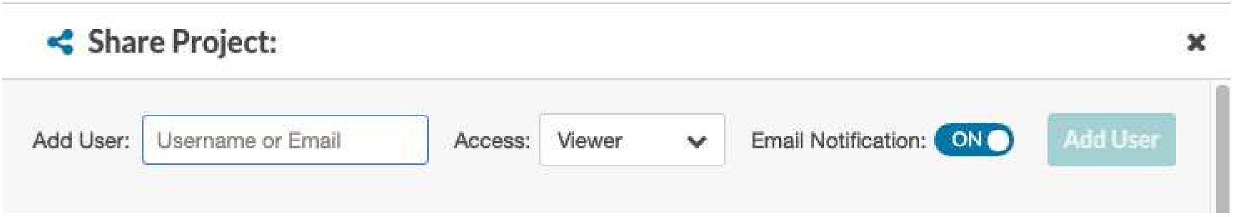
Individuals with project **admin** privileges are able to share cohorts with other DNAnexus users by granting their user account access to the project:

1. If you have Admin privileges on the project, open the project page in a web browser.
2. Click on the 👥 icon in the upper-right corner:



Access: Admin

3. Type in the username of a colleague who does not already have access to this project.
4. Select **Viewer** access



Share Project:

Add User: Username or Email

Access: Viewer

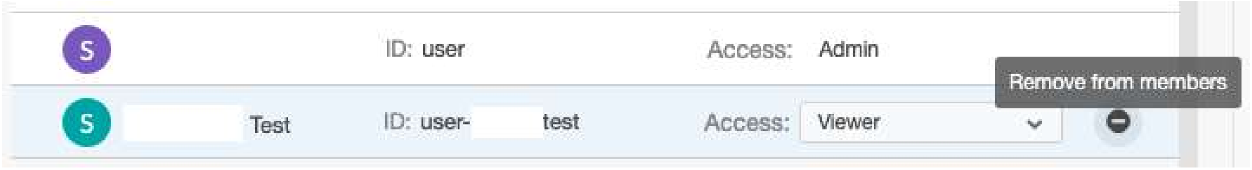
Email Notification: ON

5. Users with Viewer access to a project have read-only access to data objects and their metadata in a data container. For additional information on project permissions and sharing, please refer to [DNAnexus documentation](#).

Now confirm that your colleague can access your project and explore the cohort.

To revoke access to your project:

1. Click the 👥 icon again.
2. Click the **Remove** ⊖ icon next to the colleague's name.



S		ID: user	Access: Admin
S	[redacted] Test	ID: user-[redacted]test	Access: Viewer

Feature checklist

I confirm that I was able to:

- ▢ Find **Non-cancer illness code, self-reported | Instance 0**, **Diabetes diagnosed by doctor | Instance 0**, **Treatment/medication code | Instance 0**, **VLDL Cholesterol | Instance 0**, **Phenylalanine | Instance 0** and **Body mass index (BMI) | Instance 0** tiles
- ▢ Filter the UKB 500k Pheno/Geno(array) dataset using self-reported illness code and BMI, utilize the Data Preview feature to confirm that the cohort contained data for patients that meet that criteria
- ▢ Export and download full Data Preview as a CSV file
- ▢ Save my cohort as a Cohort object
- ▢ Open my saved cohort and see that filtering criteria and layout were exactly the same as it was when I saved my cohort

Signature 3

I, a HDRUK Apollo Early Access Evaluator, have performed the actions described above and confirm the features were available and tested by myself. For the scenario and the features described I:

▢ Accept	▢ Conditionally Accept (conditional criteria below)	▢ Reject
Signature:		Date:
Name (written):		

Conditional Acceptance:

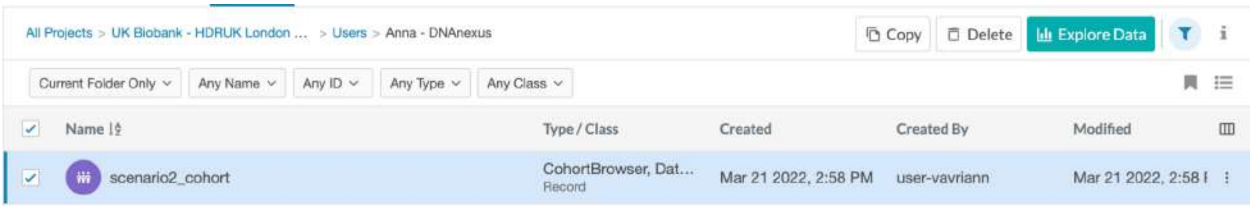
Scenario 3. Explore gene mutations

Objective

In this scenario, we will explore the genotype data from the cohort from Scenario 2. We will first investigate any High Impact mutations in **TCF7L2** gene, and export that genotype data for further review. We will do the same for Missense variations in the **CDKAL1** gene. Then we will filter the cohort down to just those participants that harbor a set of Missense mutations. We will also try to filter the cohort by specific variant ID.

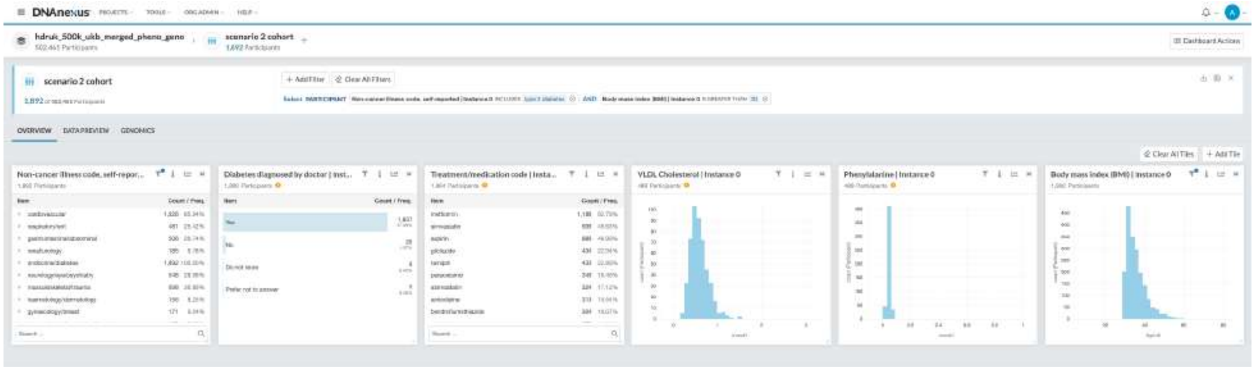
Load saved cohort

1. Return to your **User folder** within the **UK Biobank - HDRUK London UAT**.
2. Select the Cohort Browser object that represents the saved cohort from Scenario 2, and click the Explore Data button in the toolbar.



☑	Name	Type / Class	Created	Created By	Modified
☑	scenario2_cohort	CohortBrowser, Dat... Record	Mar 21 2022, 2:58 PM	user-vavriann	Mar 21 2022, 2:58 I

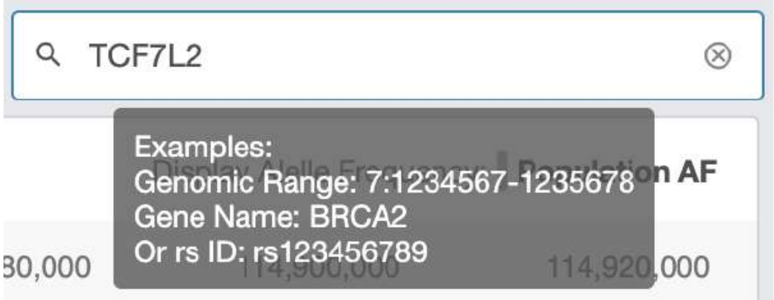
3. Confirm that you have **1,892 participants** in your cohort who meet the criteria of having BMI greater than 30 and self-reported themself as type 2 diabetic.


DNAnexus
hdruk_500k_ukb_merged_pheno_geno
scenario 2 cohort
Dashboard Actions
scenario 2 cohort
1,892
+ Add Filter
Clear All Filters
OVERVIEW
DATA PREVIEW
GENOMICS
Clear All Tiles
+ Add Tile
Non-cancer illness code, self-repor...
Diabetes diagnosed by doctor | Inst...
Treatment/medication code | Insta...
VLDL Cholesterol | Instance 0
Phenylalanine | Instance 0
Body mass index (BMI) | Instance 0

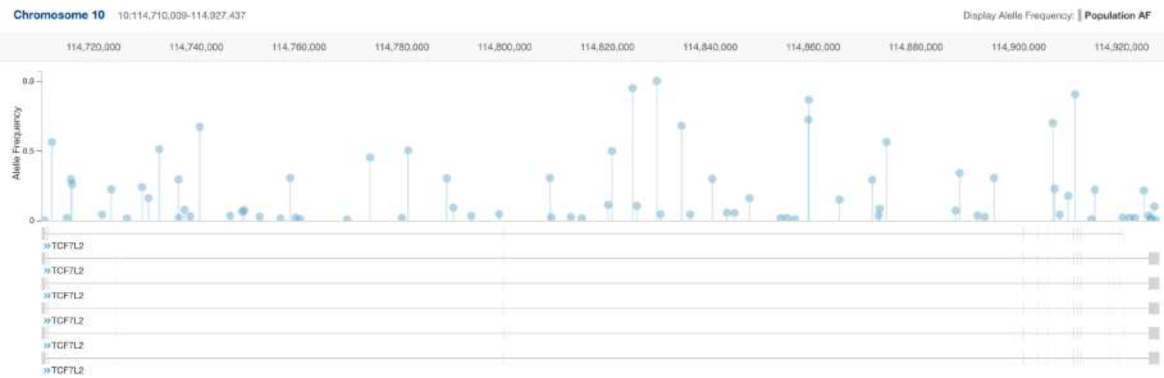
Explore TCF7L2 mutations

Now let's use the built-in Genomics browser and explore mutations in key genes implicated in diabetes phenotypes, starting with the Variant Browser.

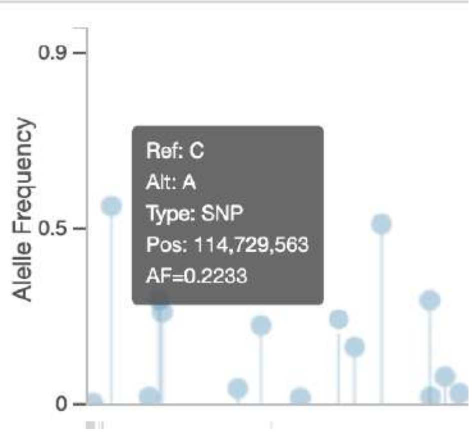
1. Navigate to the **GENOMICS** tab to the right of the **OVERVIEW** (dashboard) tab. A window with a **lollipop plot** and **Allele Table** will appear. By default, this plot is pre-populated with variants from interval on the ACOT11 gene. Change the view by typing **TCF7L2** into the search window and hit Return.


TCF7L2

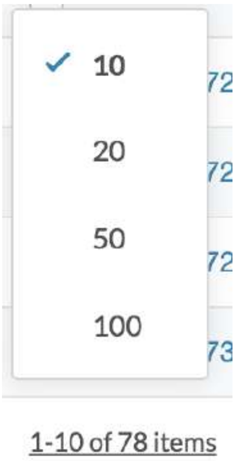
Examples:
Genomic Range: 7:1234567-1235678
Gene Name: BRCA2
Or rs ID: rs123456789



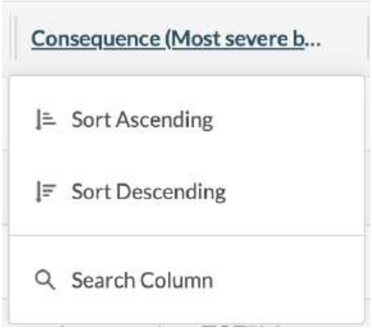
The browser has navigated to chromosome 10 and is displaying the region 10:114,710,009-114,927,437 where the gene **TCF7L2** gene resides. Below and along the X axis, it will display alternative transcripts of the gene; the Y axis plots allele frequencies of the whole dataset (population). In the plot area, it will display a large number of mutations as a lollipop plot of allele frequencies. Hovering over blue circles displays a tooltip with details for a particular variant.


Ref: C
Alt: A
Type: SNP
Pos: 114,729,563
AF=0.2233

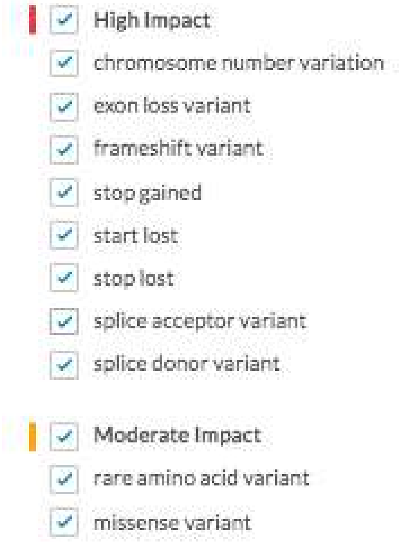
2. Scroll down to the **ALLELE TABLE** to reveal the full complement of variants for this gene for this cohort.
3. Use the link at the bottom left of the table to increase the number of variants displayed.



4. Use the Consequence column to find High or Moderate impact variants in **TCF7L2** for this cohort by selecting the **Search Column** feature. Click on the Consequence (Most severe by Gene) link in the column header to reveal the search options.

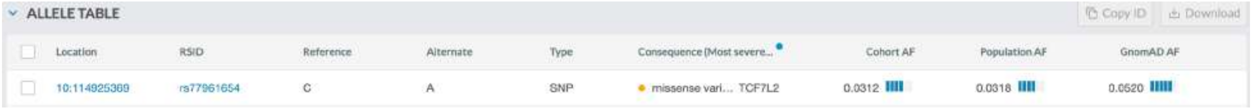


5. Click Search and then select **High Impact** and **Moderate Impact**.



- ☑ High Impact
  - ☑ chromosome number variation
  - ☑ exon loss variant
  - ☑ frameshift variant
  - ☑ stop gained
  - ☑ start lost
  - ☑ stop lost
  - ☑ splice acceptor variant
  - ☑ splice donor variant
- ☑ Moderate Impact
  - ☑ rare amino acid variant
  - ☑ missense variant

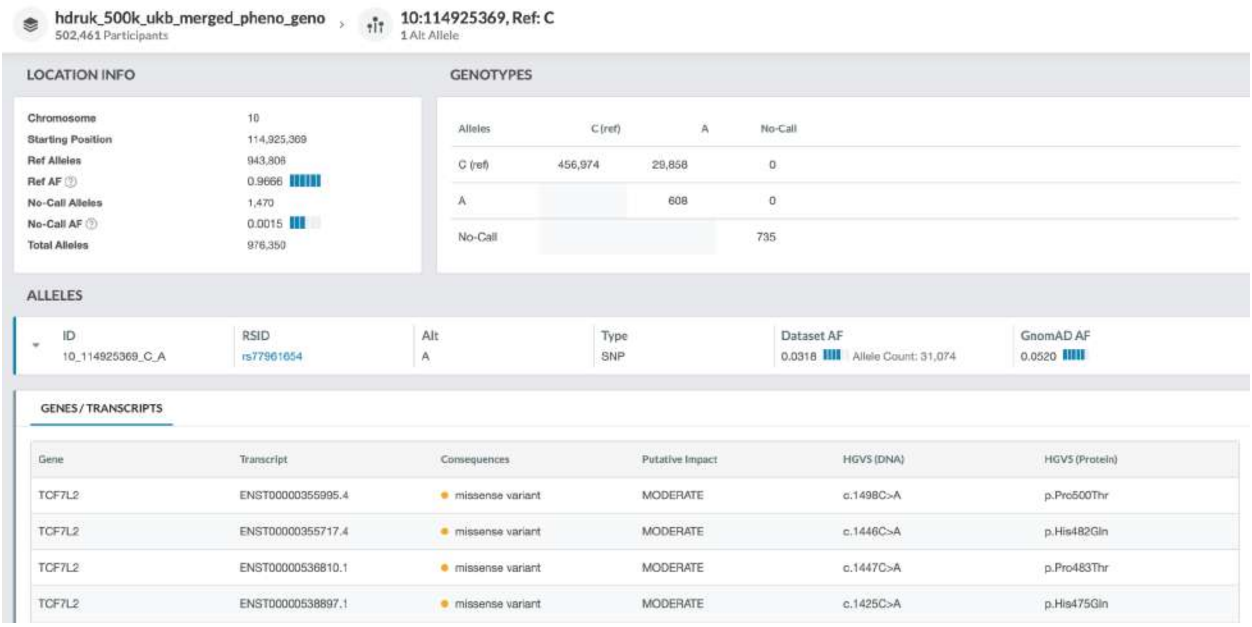
6. The filter is dynamic, returning only the variants that are High or Moderate impact for **TCF7L2** for the individuals in the filtered cohort.



ALLELE TABLE

□	Location	RSID	Reference	Alternate	Type	Consequence (Most severe...	Cohort AF	Population AF	GnomAD AF
□	10:114925369	rs77961654	C	A	SNP	● missense vari... TCF7L2	0.0312	0.0318	0.0520

7. To learn more about the variants in the table Click on any of the links in the **Location** column. A new browser tab will open and display a Variant Details page containing information about this particular variant's genotype, allele frequencies, both DNA and protein changes.



hdruk_500k_ukb_merged_pheno_geno
502,461 Participants

10:114925369, Ref: C
1 Alt Allele

LOCATION INFO

Chromosome	10
Starting Position	114,925,369
Ref Alleles	943,808
Ref AF	0.9666
No-Call Alleles	1,470
No-Call AF	0.0015
Total Alleles	976,350

GENOTYPES

Alleles	C (ref)	A	No-Call
C (ref)	456,974	29,858	0
A		608	0
No-Call			735

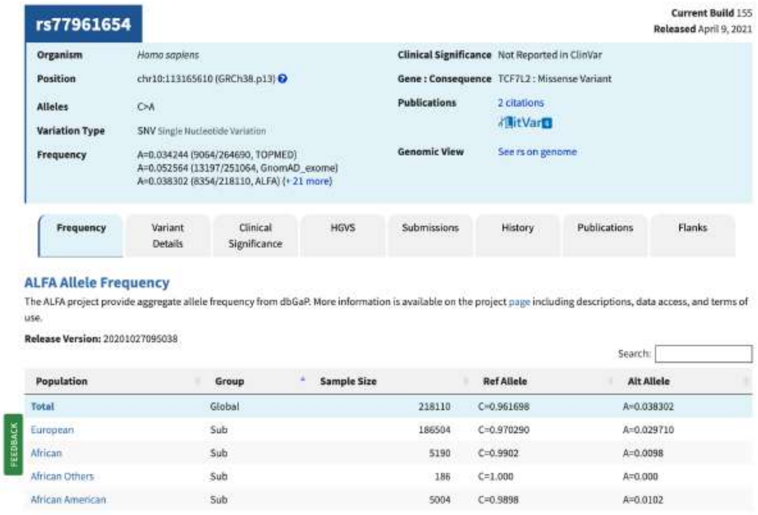
ALLELES

ID	RSID	Alt	Type	Dataset AF	GnomAD AF
10_114925369_C_A	rs77961654	A	SNP	0.0318 Allele Count: 31,074	0.0520

GENES / TRANSCRIPTS

Gene	Transcript	Consequences	Putative Impact	HGVS (DNA)	HGVS (Protein)
TCF7L2	ENST00000355995.4	● missense variant	MODERATE	c.1498C>A	p.Pro500Thr
TCF7L2	ENST00000355717.4	● missense variant	MODERATE	c.1446C>A	p.His482Gln
TCF7L2	ENST00000536810.1	● missense variant	MODERATE	c.1447C>A	p.Pro483Thr
TCF7L2	ENST00000538897.1	● missense variant	MODERATE	c.1425C>A	p.His475Gln

8. Return to the **ALLELE TABLE** in the Cohort Browser, Click on any of the **rs####** links in the **RSID** column. The new browser tab will take you to NCBI's dbSNP webpage for that variant.



rs77961654

Current Build 155
Released April 9, 2021

Organism	*Homo sapiens*	Clinical Significance	Not Reported in ClinVar
Position	chr10:113165610 (GRCh38.p13)	Gene : Consequence	TCF7L2 : Missense Variant
Alleles	C>A	Publications	2 citations LitVar
Variation Type	SNV Single Nucleotide Variation		

Frequency	A=0.034244 (9064/264690, TOPMED)
	A=0.052564 (13197/251064, GnomAD_exome)
	A=0.038302 (8354/218110, ALFA) (+ 21 more)
		Genomic View	See rs on genome

ALFA Allele Frequency

The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project page including descriptions, data access, and terms of use.

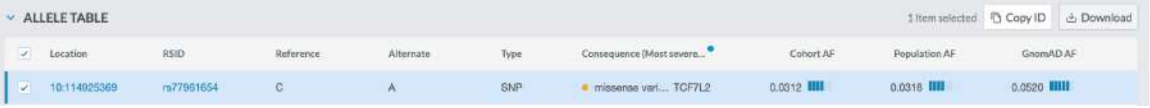
Release Version: 20201027095038

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	218110	C=0.961698	A=0.038302
European	Sub	186504	C=0.970290	A=0.029710
African	Sub	5190	C=0.9902	A=0.0098
African Others	Sub	186	C=1.000	A=0.000
African American	Sub	5004	C=0.9898	A=0.0102

Export variants table

Often, scientists would like to export variant data and explore them in third party applications. To export the variants in the **ALLELE TABLE** to a file:

1. Click on the **checkbox** on the left of the Location header. All rows will be selected.



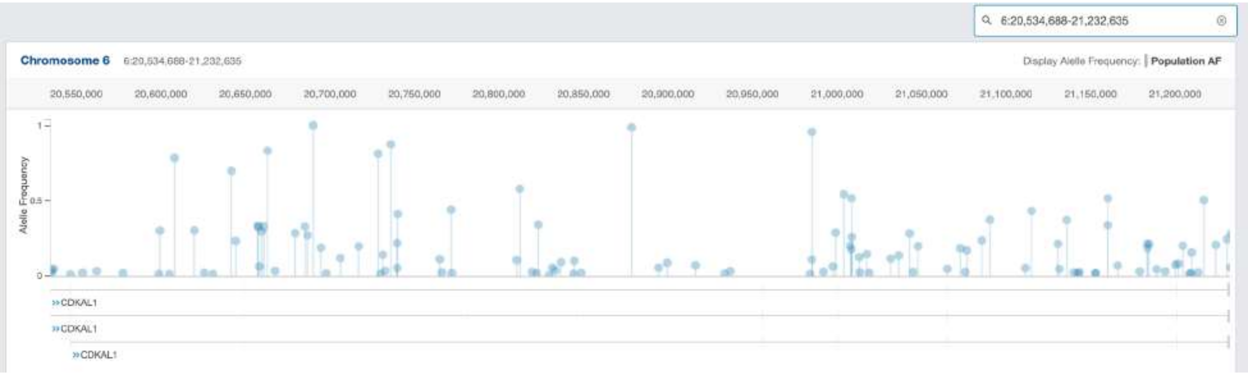
ALLELE TABLE — 1 item selected | Copy ID | Download

☑	Location	RSID	Reference	Alternate	Type	Consequence (Most severe...	Cohort AF	Population AF	GnomAD AF
☑	10:114925369	rs77961654	C	A	SNP	● missense vari... TCF7L2	0.0312	0.0318	0.0520

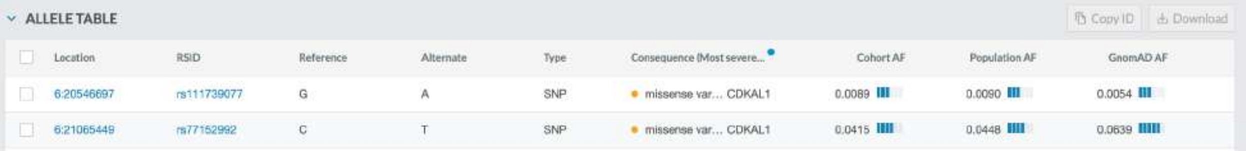
2. Click on the **Download** button in the upper right corner of the table. This table of variants will be downloaded to your computer as a csv file and can be used in third party spreadsheet applications.

We will now repeat the gene-specific exploration process and explore variants for 1 additional diabetes-related gene of interest: **CDKAL1**.

3. Go to the search box at the upper right corner of the Lollipop plot, click ⊗ to clear input, type **CDKAL1** and hit return. The lollipop plot now displays the genomic region for the **CDKAL1** gene on chromosome 6.



4. Use the Consequence column in the ALLELE TABLE to find **Missense** variants in **CDKAL1** for this cohort.



ALLELE TABLE — Copy ID | Download

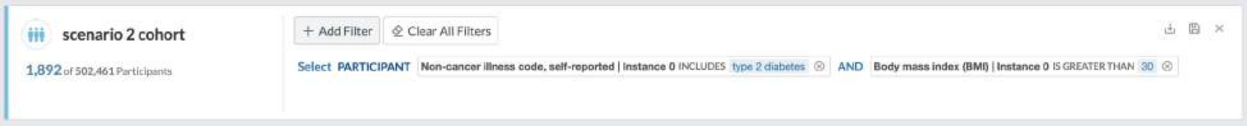
☐	Location	RSID	Reference	Alternate	Type	Consequence (Most severe...	Cohort AF	Population AF	GnomAD AF
☐	6:20546697	rs111739077	G	A	SNP	● missense var... CDKAL1	0.0089	0.0090	0.0054
☐	6:21065449	rs77152992	C	T	SNP	● missense var... CDKAL1	0.0415	0.0448	0.0639

After completing this exploration you will have identified a number of High or Moderate Impact variants in **TCF7L2** and the missense variant for **CDKAL1** gene for cohort from scenario 2. We will now use the GENO filter feature in the Cohort Browser to further refine the Cohort to only those participants who harbor specific variants.

Filter cohort to participants with Missense mutation in CDKAL1 gene

As a next step we will use the GENO filter feature to filter the cohort to just those individuals who harbor a missense variant in the CDKAL1 gene we have been exploring in previous steps using the Variant Browser.

1. Navigate back to the **OVERVIEW** (dashboard) tab and Click Add Filter

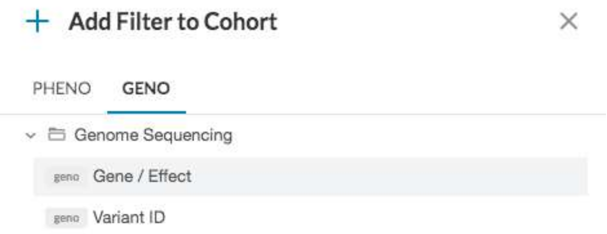


scenario 2 cohort

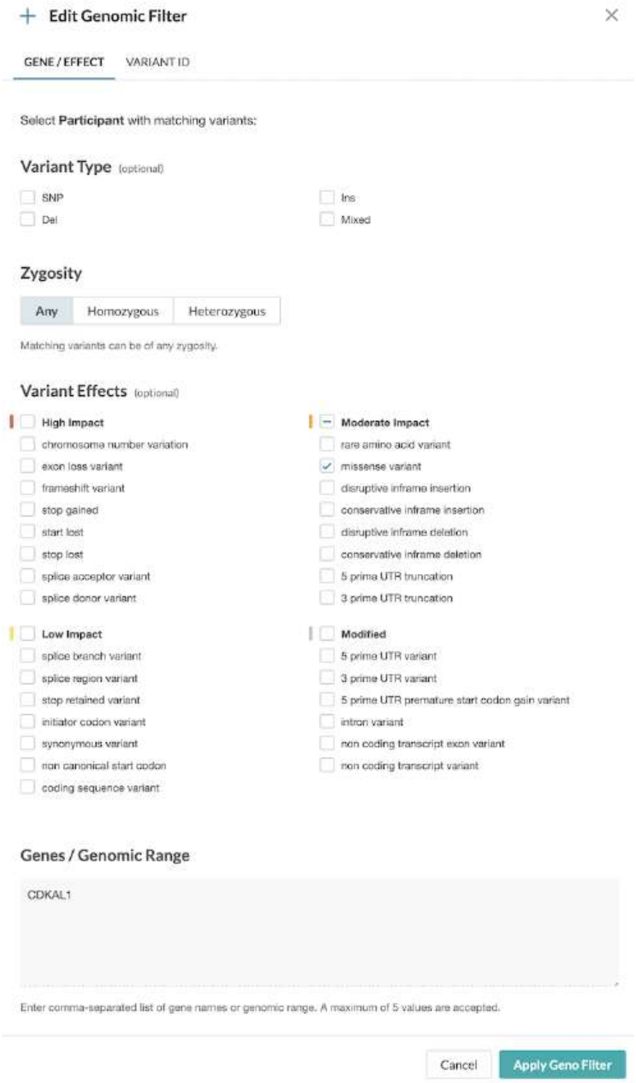
1,892 of 502,461 Participants

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes AND Body mass index (BMI) | Instance 0 IS GREATER THAN 30

2. Select **GENO** tab from the Add Filter menu and then select **Gene/Effect** to open the **Gene/Effect** filter menu.



3. Select **Missense variant** (Moderate impact) in the Variant Effects section. Type **CDKAL1** into the Genes/Genomic Range box.



+ Edit Genomic Filter ×

GENE / EFFECT | VARIANT ID

Select **Participant** with matching variants:

Variant Type (optional)

☐ SNP
☐ Del
☐ Ins
☐ Mixed

Zygosity

[✓ Any] | Homozygous | Heterozygous

Matching variants can be of any zygosity.

Variant Effects (optional)

☐ **High Impact**
☐ chromosome number variation
☐ exon loss variant
☐ frameshift variant
☐ stop gained
☐ start lost
☐ stop lost
☐ splice acceptor variant
☐ splice donor variant

☐ **Low Impact**
☐ splice branch variant
☐ splice region variant
☐ stop retained variant
☐ initiator codon variant
☐ synonymous variant
☐ non canonical start codon
☐ coding sequence variant

⊟ **Moderate Impact**
☐ rare amino acid variant
☑ missense variant
☐ disruptive inframe insertion
☐ conservative inframe insertion
☐ disruptive inframe deletion
☐ conservative inframe deletion
☐ 5 prime UTR truncation
☐ 3 prime UTR truncation

☐ **Modified**
☐ 5 prime UTR variant
☐ 3 prime UTR variant
☐ 5 prime UTR premature start codon gain variant
☐ intron variant
☐ non coding transcript exon variant
☐ non coding transcript variant

Genes / Genomic Range

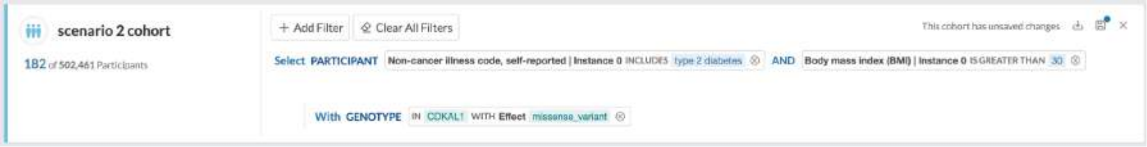
CDKAL1

Enter comma-separated list of gene names or genomic range. A maximum of 5 values are accepted.

Cancel | Apply Geno Filter

**NOTE: Gene names must be in ALL CAPS and comma separated. A maximum of 5 genes can be filtered on at one time.*

4. Click **Apply Geno Filter**. The cohort filter bar will show the newly added genomic filter:



scenario 2 cohort
182 of 502,461 Participants

+ Add Filter | Clear All Filters | This cohort has unsaved changes

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes AND Body mass index (BMI) | Instance 0 IS GREATER THAN 30

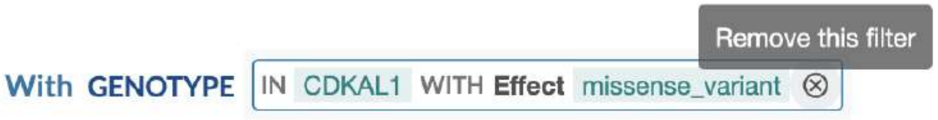
With GENOTYPE IN CDKAL1 WITH Effect missense_variant

When the filter step has completed, your cohort will contain only the participants from the **scenario 2 cohort** with missense variants in **CDKAL1** gene.

NOTE: The **Filter by gene and variant effect allows you to specify genes of interest, then select the transcribed variant effects to retain in the filter. Currently, a **maximum of 5 genes** can be entered.*

NOTE: **IMPORTANT Entering multiple genes or variants will return individuals with variants in **any** of the genes listed **OR** with any of the variants listed in the filter. The filter **will not require** individuals to have all the variants listed.*

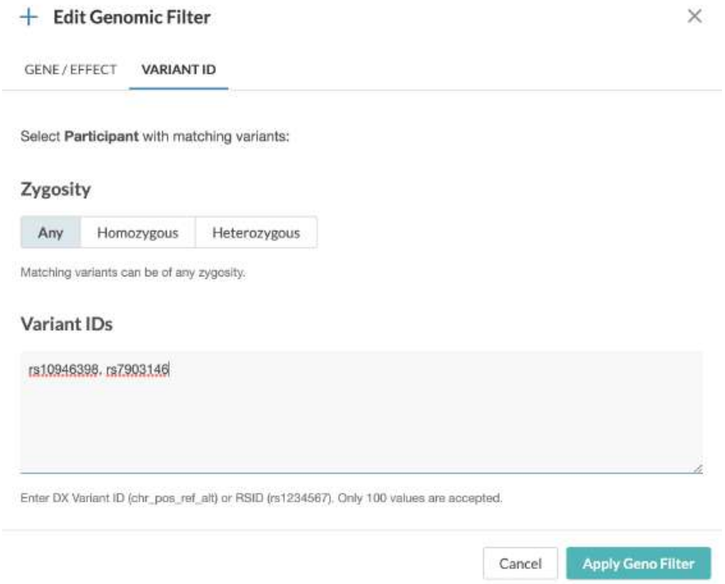
5. Now we will try to filter the dataset by variant ID. Delete the genomic filter that we added previously by clicking the x icon next to the circle.



Remove this filter

With GENOTYPE IN CDKAL1 WITH Effect missense_variant ⊗

6. Click **+ Add Filter** button and navigate to the **GENO** tab again, there choose **Variant ID**.
7. Copy and paste these rsIDs into the text field: **rs10946398**, **rs7903146**. Click **Apply Geno Filter**.



+ Edit Genomic Filter ×

GENE / EFFECT | VARIANT ID

Select **Participant** with matching variants:

Zygosity

Any | Homozygous | Heterozygous

Matching variants can be of any zygosity.

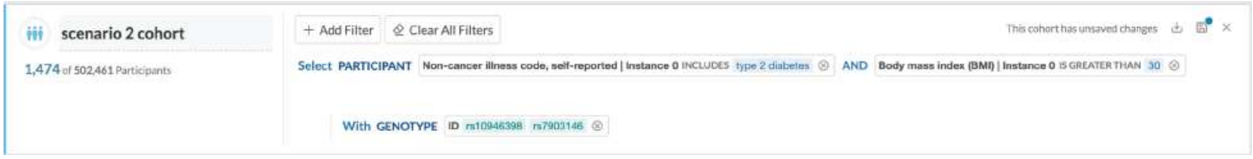
Variant IDs

rs10946398, rs7903146

Enter DX Variant ID (chr_pos_ref_alt) or RSID (rs1234567). Only 100 values are accepted.

Cancel | Apply Geno Filter

Resulting cohort should look like this.



scenario 2 cohort

1,474 of 502,461 Participants

+ Add Filter | Clear All Filters

This cohort has unsaved changes

Select PARTICIPANT Non-cancer illness code, self-reported | Instance 0 INCLUDES type 2 diabetes ⊗ AND Body mass index (BMI) | Instance 0 IS GREATER THAN 30 ⊗

With GENOTYPE ID rs10946398 rs7903146 ⊗

**NOTE: A maximum of 100 variants can be entered when filtering by variant ID.*

**NOTE: Variant ID is structured as: CHROMOSOME_POSITION_REF_ALT. When you copy variant IDs from the allele table you will get them in this format, so you can filter by them right away.*

Feature checklist

I confirm that I was able to:

- ▢ Launch a previously saved cohort in the Cohort Browser
- ▢ Explore variants in the **TCF7L2** and **CDKAL1** genes using Variant Browser features: lollipop plot, tooltips, and Allele Table
- ▢ Filter variants in the Allele Table using the Consequences column
- ▢ View specific gene variants in the Allele Table, including link outs to Allele Details
- ▢ Download a CSV table with data about selected variants.
- ▢ Apply a genomic filter to identify participants in the cohort with missense variation in CDKAL1 and apply a genomic filter by variant ID (rsID)

Signature 4

I, a HDRUK Apollo Early Access Evaluator, have performed the actions described above and confirm the features were available and tested by myself. For the scenario and the features described I:

▢ Accept	▢ Conditionally Accept (conditional criteria below)	▢ Reject
Signature:		Date:
Name (written):		

Conditional Acceptance:

Scenario 4. Comparing Cohorts with different Genotype and Demographic attributes

Objective

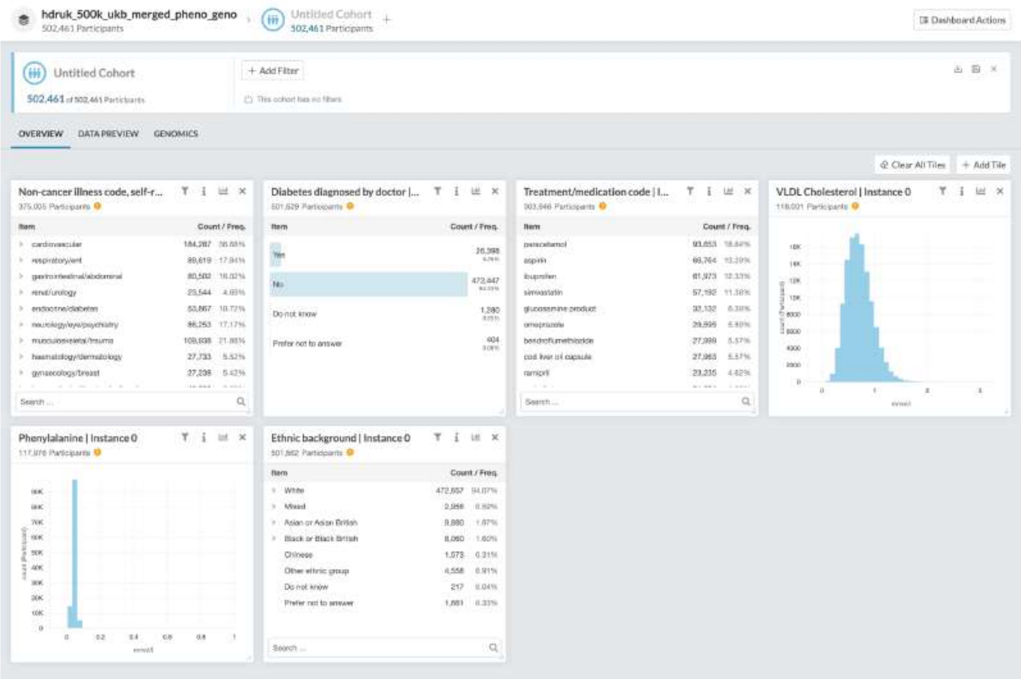
In this scenario, we will explore differences in the cohorts of diverse ethinc background with High or Moderate Impact mutations in **PAX4** gene. How is the distribution of type 2 diabetic patients in both cohorts? What do the VLDL cholesterol, phenylalanine distributions look like for each cohort? Can you spot any differences in used treatment?

Create two Cohorts that differ by ethnic background

For this scenario we will start from the entire unfiltered UK Biobank cohort.

1. Open dashboard we saved in scenario 2 – **test_dashboard**. Remove **Body mass index (BMI) | Instance 0** tile and add tile for **Ethnic background | Instance 0**.
2. You should now see an unfiltered and untitled cohort with following tiles:
    a. **Non-cancer illness code, self-reported | Instance 0**
    b. **Diabetes diagnosed by doctor | Instance 0**
    c. **Treatment/medication code | Instance 0**
    d. **Ethnic background | Instance 0**
    e. **VLDL Cholesterol | Instance 0**
    f. **Phenylalanine | Instance 0**

The resulting dashboard should look similar to the image below.



Non-cancer illness code, self-r...
375,005 Participants

Item	Count / Freq.	
cardiovascular	184,287	36.68%
respiratory/ent	89,619	17.84%
gastrointestinal/abdominal	80,502	16.02%
renal/urology	23,544	4.69%
endocrine/diabetes	53,867	10.72%
neurology/eye/psychiatry	86,253	17.17%
musculoskeletal/trauma	109,938	21.88%
haematology/dermatology	27,733	5.52%
gynaecology/breast	27,236	5.42%

Diabetes diagnosed by doctor |...
501,629 Participants

Item	Count / Freq.	
Yes	26,398	5.26%
No	472,447	94.03%
Do not know	1,280	0.25%
Prefer not to answer	404	0.08%

Treatment/medication code | I...
503,946 Participants

Item	Count / Freq.	
paracetamol	93,653	18.64%
aspirin	66,764	13.29%
ibuprofen	61,973	12.33%
simvastatin	57,192	11.38%
glucosamine product	32,132	6.39%
omeprazole	29,695	5.91%
bendroflumethiazide	27,999	5.57%
cod liver oil capsule	27,963	5.57%
ramipril	23,235	4.62%

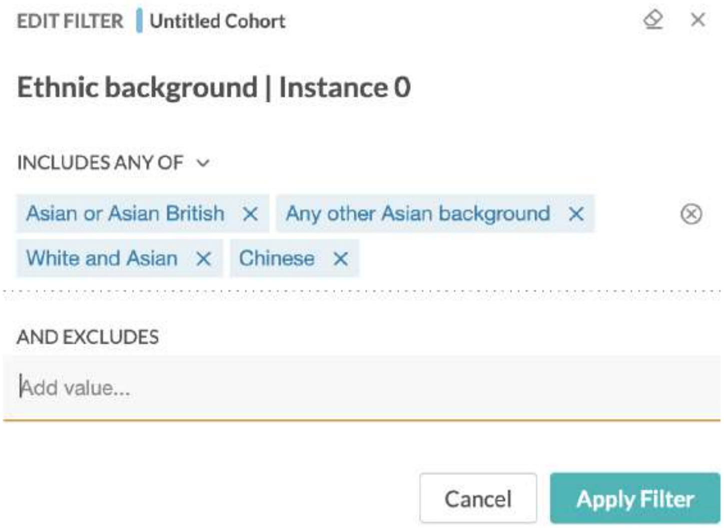
VLDL Cholesterol | Instance 0
118,001 Participants

Phenylalanine | Instance 0
117,976 Participants

Ethnic background | Instance 0
501,662 Participants

Item	Count / Freq.	
White	472,657	94.07%
Mixed	2,956	0.59%
Asian or Asian British	9,880	1.97%
Black or Black British	8,060	1.60%
Chinese	1,573	0.31%
Other ethnic group	4,558	0.91%
Do not know	217	0.04%
Prefer not to answer	1,661	0.33%

3. Add phenotypic filter for **Ethnic background | Instance 0** filter for values (INCLUDES ANY OF): **Asian or Asian British**, **Any other Asian background**, **White and Asian**, **Chinese**.


EDIT FILTER | Untitled Cohort

Ethnic background | Instance 0

INCLUDES ANY OF

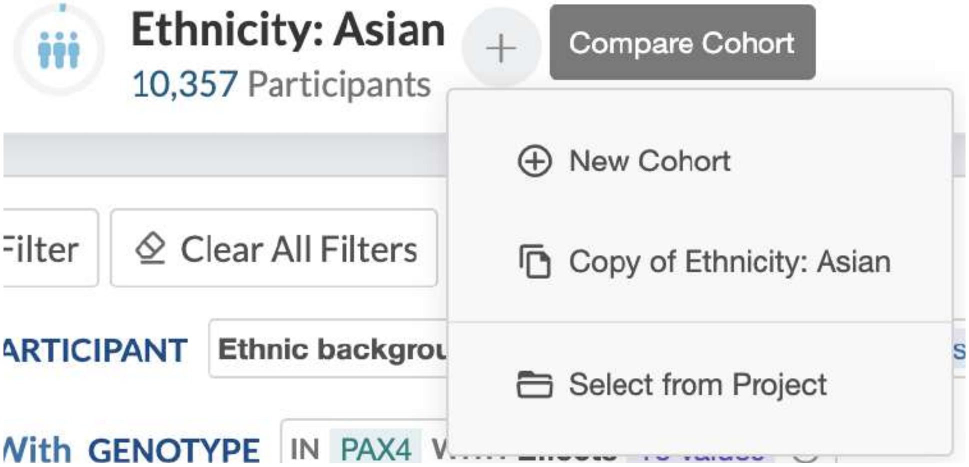
Asian or Asian British ✕ Any other Asian background ✕ White and Asian ✕ Chinese ✕

AND EXCLUDES

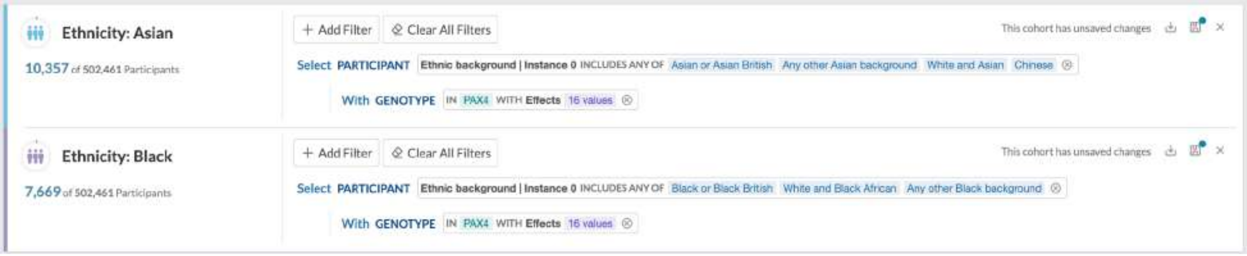
Add value...

4. Use the **Add Filter**, **Geno** filter for participants that have High or Moderate Impact mutations in PAX4 gene.
5. When the filter completes, edit the name of this new cohort directly in the name field.

6. Create a comparator cohort by copying the cohort we just created.



7. Adjust the cohort filter for the **Copy of Ethnicity: Asian** cohort, you can do it by clicking the **Ethnic background | Instance 0** part of the cohort filter and setting it to include following values: **Black or Black British, White and Black African**, **Any other Black background**. Keep the same genomic filter.
8. Change the name of the cohort by editing the name directly. Your cohort definitions should look like this:



Ethnicity: Asian
10,357 of 502,461 Participants
Select PARTICIPANT Ethnic background | Instance 0 INCLUDES ANY OF Asian or Asian British Any other Asian background White and Asian Chinese
With GENOTYPE IN PAX4 WITH Effects 16 values

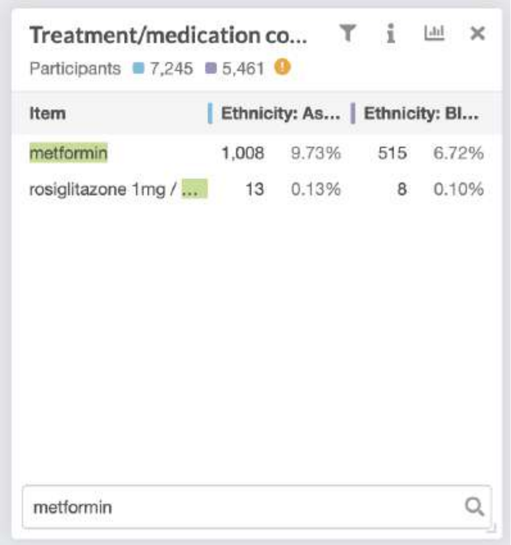
Ethnicity: Black
7,669 of 502,461 Participants
Select PARTICIPANT Ethnic background | Instance 0 INCLUDES ANY OF Black or Black British White and Black African Any other Black background
With GENOTYPE IN PAX4 WITH Effects 16 values

Your **Cohort Comparison** dashboard should now show values for participants with Ethnic Background = Asian and with High/Moderate Variants in PAX4 gene (Blue) compared side by side with values for Ethnic Background = Black and with High/Moderate Variants in PAX4 gene (Purple).

Compare the Distribution of Values from Multiple Fields between the 2 Cohorts

We can now use the collection of fields in the Dashboard to compare values for various fields/attributes between the two cohorts we created.

1. Review the values in the **Treatment/medication code | Instance 0** more closely by searching for the term **metformin**.



Treatment/medication co...

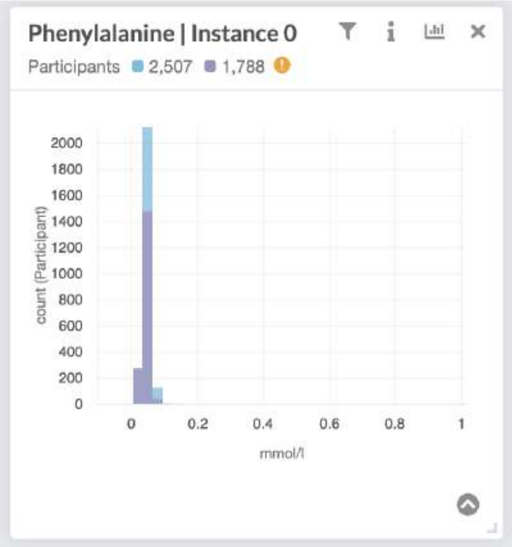
Participants ■ 7,245 ■ 5,461

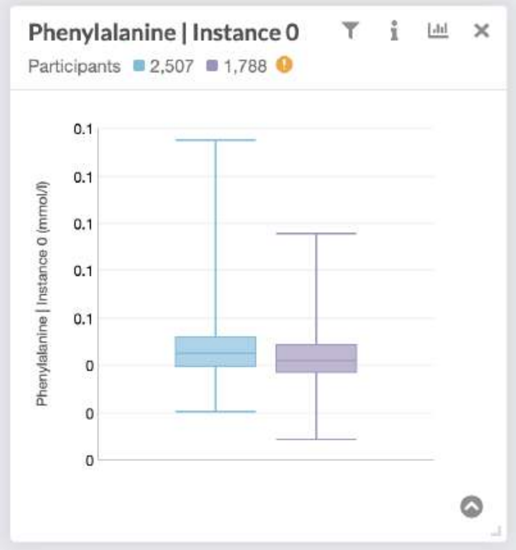
Item	Ethnicity: As...		Ethnicity: Bl...	
metformin	1,008	9.73%	515	6.72%
rosiglitazone 1mg / ...	13	0.13%	8	0.10%

metformin

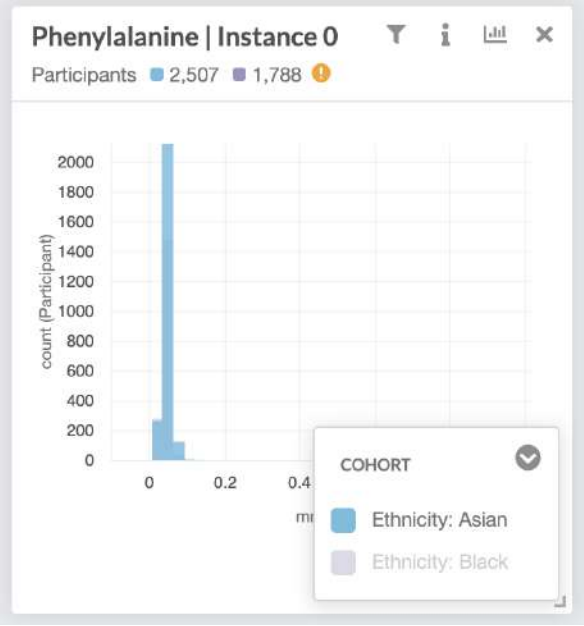
While the sample sizes are small, there may be significant differences in the % of participants who use metformin of Asian ethnicity (9.73%) compared to the % of participants of Black ethnicity (6.72%).

To be able to more closely explore any differences in **Phenylalanine | Instance 0** measurements between our cohorts, change the **Chart Type** to Box Plot. Hovering over a box plot will reveal the Min, Med, Max values, which facilitates easy comparison between cohorts.



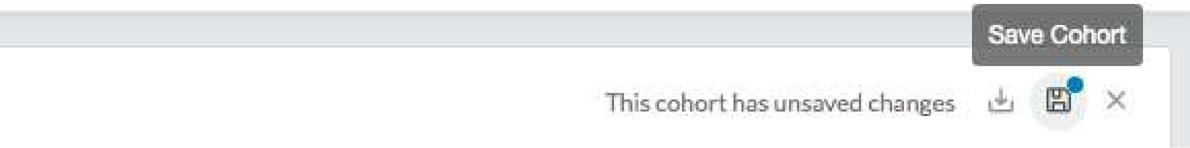


To view histograms for each cohort separately, you can use the ⌃ icon and choose the cohort of your interest.



Save the Case and Comparator Cohorts to your Project folder

1. To **save** these comparator cohorts that we created during this Cohort Compare workflow, click on the Save icon in the top right section of each Cohort description. Be sure to save each cohort (if desired) for future exploration.



Feature checklist

I confirm that I was able to:

- ▢ Create and Compare 2 cohorts that differ by Ethnicity
- ▢ Compare Cohorts across multiple categorical and numerical values using the collection of demographic and clinical data elements in the Cohort Dashboard.
- ▢ Save the Case and Comparator Cohorts in my personal User folder within the Project.

Signature 5

I, a HDRUK Apollo Early Access Evaluator, have performed the actions described above and confirm the features were available and tested by myself. For the scenario and the features described I:

▢ Accept	▢ Conditionally Accept (conditional criteria below)	▢ Reject
Signature:		Date:
Name (written):		

Conditional Acceptance: