Summary of the number of cases for the Confluence

Updated on 11th July 2025

The size of the full cohort is **513228**, which includes male and female.

The endpoints data contains information of breast cancer (ICD10 C50; diy_p1232015483), Carcinoma in situ of breast (D05; diy_p1781594777), and other types of cancer (diy_p1176798249). Note that the total number of rows in "endpoints" file for the full cohort is **513228**.

The total number of participants with genotype data is **100706**, including **57661** women and **43045** men. We primarily focus on female participants in this study.

(i) the full cohort (513228) – prevalent breast cancer
we can count the number of breast cancer prevalence with ((df['cancer_diag'] == 1) & (df['cancer_Breast'] == 1)):

- For women, the count is **567**.
- For men, the count is **1**.

(ii) the participants with genetic data (100706) – prevalent breast cancer
we can count the number of breast cancer prevalence with ((df['cancer_diag'] == 1) & (df['cancer_Breast'] == 1)):

- For women, the count is **107**.
- For men, the count is **0**.

(iii) the full cohort (513228) with breast cancer endpoints
breast cancer (ICD10 C50; ep_diy_p1232015483_combined_ep == 1):

- For women, the count is **3113**.
- For men, the count is **57**.

carcinoma in situ breast (D05; ep_diy_p1781594777_combined_ep == 1):

- For women, the count is **81**.
- For men, the count is **4**.

For women, the overlap between C50 and D05 is **67**.
For men, the overlap between C50 and D05 is **1**.

(iv) the participants with genetic data (100706) with breast cancer endpoints
breast cancer (ICD10 C50; ep_diy_p1232015483_combined_ep == 1):

- For women, the count is **574**.
- For men, the count is **12**.

carcinoma in situ of breast (D05; ep_diy_p1781594777_combined_ep == 1):

- For women, the count is **14**.
- For men, the count is **1**.

For women, the overlap between C50 and D05 is **11**.
For men, the overlap between C50 and D05 is **0**.

(v) the population-based subset of the genetics (77176) with breast cancer endpoints
breast cancer (ICD10 C50; ep_diy_p1232015483_combined_ep == 1 &
is_in_gwas_population_subset == 1):

- For women, the count is **510**.
- For men, the count is **7**.

carcinoma in situ of breast (D05; ep_diy_p1781594777_combined_ep == 1 &
is_in_gwas_population_subset == 1):

- For women, the count is **12**.
- For men, the count is **1**.

For women, the overlap between C50 and D05 is **9**.
For men, the overlap between C50 and D05 is **0**.

(vi) case-control cohort for the Confluence (women only) after exclusion for prior non-breast cancer
There are **57661** women from 100706 partcipants with genetic data.
The number of controls is **42765** after excluding non-population subsets and incident cases of other cancer to ensure controls are individuals with no history of invasive cancer; we also exclude cases that developed **other cancer before breast cancer**:

- **Overall breast cancer** (ICD10 C50 or D05 or breast cancer prevalence ==1), the count is **551.**

- **ER negative breast cancer** (has subtype of ER negative based on the overall breast cancer), the count is **48**.
- **ER positive breast cancer** (has subtype of ER positive based on the overall breast cancer), the count is **113**.
- **Triple negative breast cancer** (has subtype of triple negative based on the overall breast cancer), the count is **13.**

Note that: in our previous summary statistics by SAIGE and REGENIE, we also considered the situation where we included the cases even if these participants are in the non-population subset. For example, the number of breast cancers cases (breast_cancer_prevalence == 1 & ep_diy_p1232015483_combined_ep == 1 & is_in_gwas_population_subset == 0) in the non-population subset is **71** in our previous analysis.