Friedrich-Alexander-Universität
Technische Fakultät

FAU

Project ‚BRFC'
Behavior Risk Factors & Cancer

Agnes Thanbichler
nilapalin

Setting the context

Setting the context

Get the idea

- Cancer among leading causes of deaths
- Behavioral risk factors for cancers rising: obesity, physical inactivity, diet
- at least 13 types of cancer related to overweight or obesity

Lets combine cancer and behavioral risk factor datasets

and show a connection!

https://www.cancer.gov/about-cancer/causes-prevention/risk/obesity/obesity-fact-sheet

Hypotheses

Hypotheses

Get the idea - be more precise

- **Behavioral risk factors**
- **are influencing the chance to develop cancer**
- **are influencing the chance of mutation(s) in genes which are related to obesity (and cancer)**

What data?

BRFSS | GDC

Open data from the U.S.

BRFSS - Behavioral Risk Factor Surveillance System

Health surveys collecting data about United States residents, run by the Centers for Disease Control and Prevention.

BRFSS FAQ

- Comma separated values file for download
- Percent of adults aged 18 years and older who have an overweight classification
- Percent of adults aged 18 years and older who have obesity
- Percent of adults who engage in muscle-strengthening activities on 2 or more days a week
- Percent of adults who engage in no leisure-time physical activity

GDC - Genomic Data Commons

Research program of the National Cancer Institute (NCI) in the United States. GDCs mission is to provide a unified repository to the cancer research community to support research to help all people to live longer, healthier lives.

GDC - About the GDC

REST and GraphQL WebAPI

Used GraphQL to only load data which is used in that report.

- year of diagnosis, age at diagnosis, mutation in specific genes
- Genes: SIM1, POMC, LEPR, MRAP2, ADCY3, NTRK2, MC4R, KSR2, LEP, PCSK1, BDNF, SH2B1

Genes contributing to obesity The genetics of obesity: from discovery to biology Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. Nat Rev Genet. 2022 Feb;23(2):120-133. doi: 10.1038/s41576-021-00414-z. Epub 2021 Sep 23. PMID: 34556834; PMCID: PMC8459824.

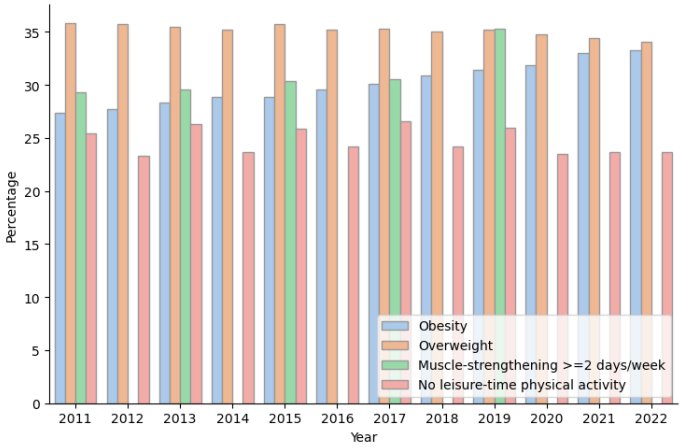
BRFSS | GDC

Data structure and Limitations

BRFSS - Behavioral Risk Factor Surveillance System

Data for aged 18 and older

Data for years 2011 to 2022



GDC - Genomic Data Commons

Submissions are on a free basis. Features may not be submitted (age, residence).

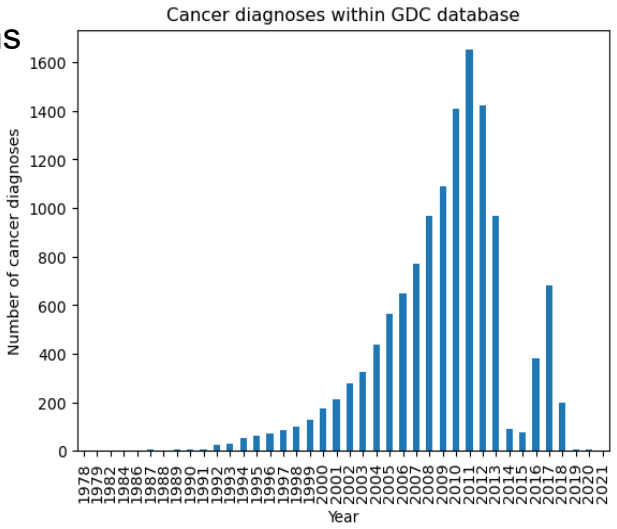
Diagnoses with residence (at enrollment) in U.S. have

- max age = 29
- number diagnoses for aged 18 and older within U.S. = 3

Drop of submissions

in year

2012/2013/2014



Limitations

Diagnoses in GDC are not restricted by residence (at enrollment) = U.S.

Only data from year 2011 is considered (present in BRFSS and GDC)

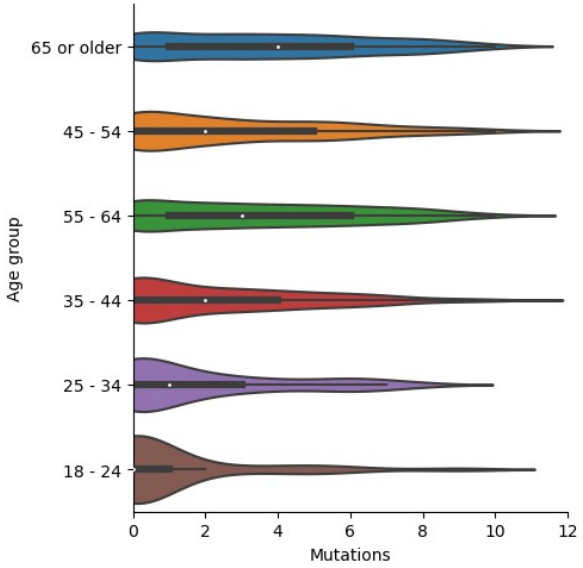
Results?

Correlation of BRFSS and Cancer

Based on age at diagnosis

age groups

BRFSS data is aggregating to age groups, therefore GDC data is also aggregated to those age groups with feature age at diagnosis.



Patients with higher age most likely have also more mutated genes (related to obesity).

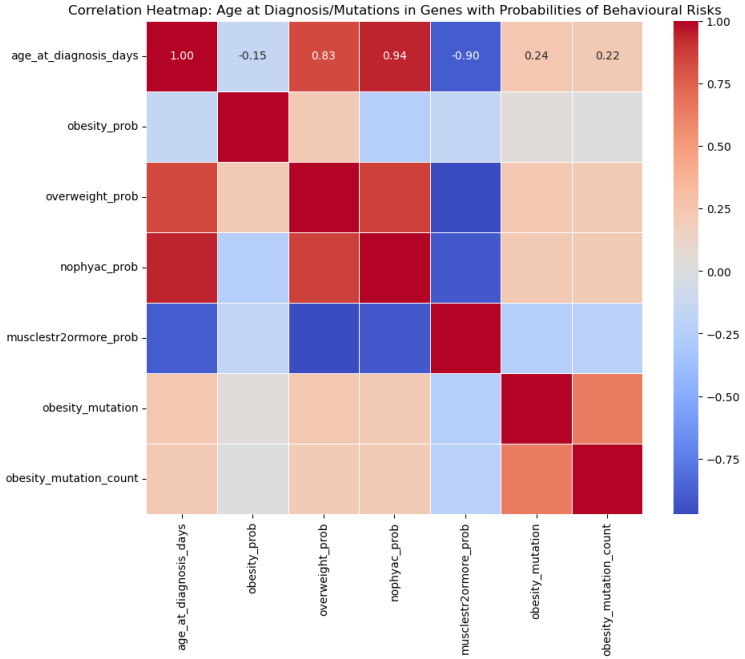
Correlation BRFSS and Age at diagnosis

The correlation matrix shows correlation between:

Age at diagnosis and

- Probability of muscle strengthening 2 or more days a week (-0.90)
- Probability of no physical activity (0.94)
- Probability of overweight (0.83)

Other combination do not show significant correlatic

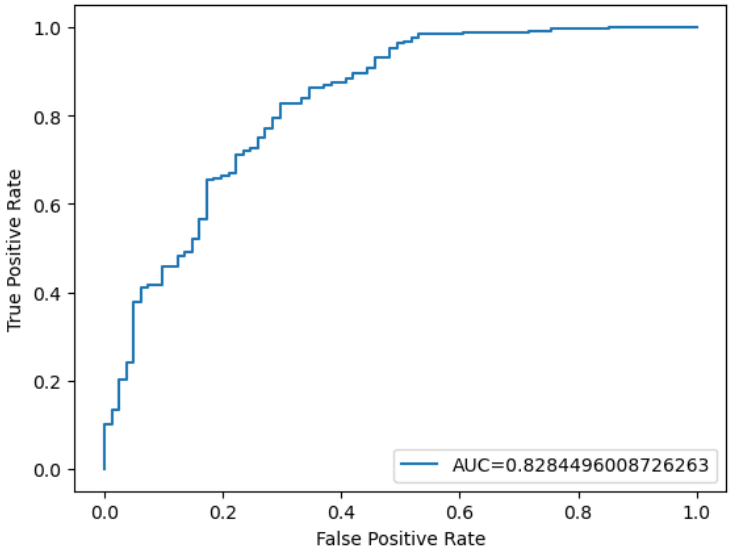


Classification

Binary classification - Logistic Regression

Is there a mutation in one or more genes given behavioral risk factors, age, primary site and disease type?

Logistic Regression classification report	Precision	Recall	F1-Score	Support
0	0.90	0.47	0.62	81
1	0.85	0.98	0.91	249
accuracy			0.86	330
macro avg	0.88	0.73	0.77	330
weighted avg	0.86	0.86	0.84	330



Good precision for both classes (0: no mutation, 1: mutation), small number in False Positives

Drop in Recall and F1-Score for class 0 (no mutation): Some are classified as class 1 (mutated) but are class 0 (not mutated)

Tend to classify more towards class 1 (mutated).

ROC (Receiver Operator Curve) and AUC (Area under Curve) ok.

Conclusion

The report shows that mutations in genes related to obesity for patients with cancer diagnosis could potentially be linked to the behavioural risk factors.

Further investigations needed and more (complete) data would be beneficial!

Thank you for your attention!