Module_3: *Cancer*

Team Members:

Cecilia and Melanie

Project Title:

A Study between the TERT Gene, Gender, and Prevalence of Cancer Types

Project Goal:

This project seeks to… *investigate the level of TERT gene in different genders and see A: are certain cancers more prone to specific levels of TERT and B: are specific levels of TERT found among different genders and therefore are certain genders more prone to a type of cancer with TERT as a metric.*

Disease Background:

- Cancer hallmark focus: Limitless replicative potential
- Overview of hallmark: Limitless replicative potential is a hallmark of cancer cells where they achieve unlimited division and replication, essentially achieving immortality. While other acquired capabilities (growth signal autonomy, insensitivity to antigrowth signals, and apoptosis resistance) break cell growth away from the environment, alone will not always lead to expansive growth because cells have intrinsic pathways that limit their growth. Cells have limited replication each generation (Hayflick limit). Replicative generations are counted by loss of telomeric DNA from ends of every chromosome during the cell cycle. The eventual loss of telomeric DNA leads to cell senescence through a DNA damage response (p53 or RB). Even if a cell bypasses senescence, the lack of telomeres protecting the ends of the chromosomes causes genomic instability and then cell death. A reactivation of telomerase happens through a ribonucleoprotein enzyme that elongate telomeres using its own RNA template (TERC) and catalytic reverse transcriptase component (TERT). This happened in 80 to 95% of human cancers and allows continued replication.
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate): TERT: TERT encodes the catalytic subunit of the enzyme telomerase, which is responsible for adding repetitive DNA sequences (TTAGGG) to the ends of chromosomes(telomeres). TERT expression is tightly controlled at the transcriptional level. In cancer, TERT promoter mutations (especially C228T and C250T) create new binding sites for transcription factors (e.g., ETS/TCF family), leading to abnormal upregulation of telomerase. Additionally, TERT can be influenced by MYC oncogene activation and Wnt/β-catenin signaling, both of which can enhance TERT transcription. This

reactivation allows cells to maintain chromosomal integrity while escaping senescence, providing a key mechanism for tumor immortality.

- Prevalence & incidence
  - Global burden (2022): ~20 million new cancer cases and 9.7 million deaths worldwide, a 0.25% rate of incidence; projections exceed 35 million new cases annually by 2050 as populations grow and age (ACS Journals+2Global Cancer Observatory+2).
  - United States: An estimated 2,041,910 new cancer cases and 618,120 deaths were reported in 2025, meaning a roughly 0.6% rate of incidence. Around 18 million Americans were alive as of 2022 and have been diagnosed with cancer, resulting in 5.4% of prevalence.
- Risk factors (genetic, lifestyle) & Societal determinants
  - Tobacco (Largest preventable driver of cancer deaths), Alcohol, Overweight/obesity, Infections (HPV, HBV/HCV, H. pylori, etc.) (IARC)
  - Environmental/occupational exposures (e.g., air pollution contributes substantially to lung adenocarcinoma burden). (The Guardian)
  - Genetic predisposition: Germline variants (e.g., BRCA1/2, MLH1/MSH2 in Lynch syndrome, APC in FAP) increase lifetime risk and often alter screening and prevention strategies (risk-reducing surgery, earlier surveillance). (General oncogenetics consensus; see NCI treatment/risk pages.) (cancer.gov)
  - Social determinants of health (SDOH): Neighborhood deprivation, housing/transportation barriers, insurance status, structural racism, language access, health literacy influence exposure, screening uptake, stage at diagnosis, treatment quality, and survival. Interventions addressing SDOH increase screening rates by a median 8.4 percentage points in U.S. studies
- Standard of care treatments (& reimbursement)
  - Core modalities (often combined, site- and stage-specific): Surgery (curative for many localized solid tumors) and Radiation therapy (definitive, adjuvant, or palliative)
  - Systemic therapy including: Chemotherapy (cytotoxic), Endocrine therapy (e.g., ER+ breast, prostate), Targeted therapy (e.g., EGFR/ALK in NSCLC; BRAF in melanoma), and Immunotherapy (checkpoint inhibitors, cytokines; cellular therapies for select hematologic malignancies) (cancer.gov+1)
  - Biomarker testing increasingly guides therapy selection (e.g., PD-L1, MSI-H/dMMR, NTRK fusions) (cancer.gov)
  - The official federal registry now has 100% coverage across all 50 American states and DC
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)
  - Anatomy & organ physiology: Cancers arise from virtually all tissues; organ context shapes biology and therapy (e.g., barriers like the blood–brain barrier in glioma; hormonal milieus in breast/prostate; unique microenvironments in liver or pancreas).

- Cell & molecular physiology: Tumorigenesis reflects accumulated genomic/epigenomic alterations plus microenvironmental changes. Canonical processes include:
  - Oncogene activation (e.g., KRAS, EGFR, MYC) and tumor-suppressor loss (TP53, RB1, APC).
  - Genomic instability (defects in DNA repair—BRCA1/2, MLH1/MSH2).
  - Telomere maintenance enabling limitless replicative potential via TERT reactivation or ALT pathways.
  - Evasion of growth suppression (RB pathway), resistance to apoptosis (p53/BCL2 family), sustained proliferative signaling, angiogenesis, invasion & metastasis, deregulated cellular energetics (Warburg effect), and immune evasion (PD-L1 upregulation, MHC loss).
  - Tumor microenvironment: stromal cells, immune cells, vasculature, and extracellular matrix shape progression and drug response (e.g., immune checkpoints targeted by ICIs). (Mechanism overview aligned with NCI treatment framework.) (cancer.gov)

Data-Set:

The dataset is derived from the The Cancer Genome Atlas (TCGA) program, which collected clinicopathologic annotation data together with multi-platform molecular profiling (genomic, transcriptomic, etc.) of human tumors. (PubMed+2NCI Genomic Data Commons+2.) Specifically, more than 11,000 human tumor samples, from 33 different cancer types are found in this dataset.

The authors created a standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR). (ScienceDirect+1). TCGA-CDR includes four major clinical outcome endpoints:

- Overall Survival (OS)
- Progression-Free Interval (PFI)
- Disease-Free Interval (DFI)
- Disease-Specific Survival (DSS)

Source of molecular data and how TERT was measured: The TCGA-CDR itself primarily contains clinical outcome data (survival, follow-up, etc.), but it is directly linked to the TCGA molecular datasets, which include:

- RNA sequencing (RNA-seq) for gene expression
- Whole-exome sequencing (WES) for somatic mutations
- Copy number variation (CNV) from SNP arrays
- DNA methylation profiling (450K or 27K arrays)
- Promoter mutation data (notably for TERT in certain cancers)

Data Analysis:

Methods

The machine learning technique we are using is: *Classification, a supervised learning method*

What is this method optimizing? How does the model decide it is "good enough"? *The method optimizes by minimizing a logistic lost function, which is then scored for accuracy using a function from the sklearn module, as well as a confusion matrix.*

Analysis

```
### Exploratory data analysis (EDA) on a cancer dataset
# Loading the files and exploring the data with pandas
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
data = pd.read_csv(
'/Users/Melanie/Downloads/Comp_BME/Module3_Cancer/GSE62944_subsample_l
og2TPM.csv', index_col=0, header=0)
metadata_df = pd.read_csv(
'/Users/Melanie/Downloads/Comp_BME/Module3_Cancer/GSE62944_metadata.cs
v', index_col=0, header=0)
print(data.head())

# Explore the data
print(data.shape)
print(data.info())
print(data.describe())

# Explore the metadata
print(metadata_df.info())
print(metadata_df.describe())

# Subset by index (genes)
desired_gene_list = ['TERT']
gene_list = [gene for gene in desired_gene_list if gene in data.index]
for gene in desired_gene_list:
    if gene not in gene_list:
        print(f"Warning: {gene} not found in the dataset.")
# .loc[] is the method to subset by index labels
# .iloc[] will subset by index position (integer location) instead
gene_data = data.loc[gene_list]
print(gene_data.head())

# Basic statistics on the subsetted data
print(gene_data.describe())
print(gene_data.var(axis=1)) # Variance of each gene across samples
# Mean expression of each gene across samples
print(gene_data.mean(axis=1))
# Median expression of each gene across samples
print(gene_data.median(axis=1))
```

```
        TCGA-E9-A1NI-01A-11R-A14D-07  TCGA-E2-A1LK-01A-21R-A14D-07  \
A1BG                        3.397369                      3.466089
```

```
A1CF                          0.008857                          0.039562
A2M                           7.575125                          6.643613
A2ML1                         0.397610                          7.625124
A4GALT                        5.277425                          5.244677

        TCGA-BH-A0B2-01A-11R-A10J-07  TCGA-E2-A107-01A-11R-A10J-07  \
A1BG                          3.789771                          3.967578
A1CF                          0.065051                          0.000000
A2M                           9.024479                          7.573842
A2ML1                         0.428689                          0.465410
A4GALT                        4.072650                          4.208381

        TCGA-LL-A5YN-01A-11R-A28M-07  TCGA-BH-A0DQ-01A-11R-A084-07  \
A1BG                          4.733007                          3.011343
A1CF                          0.014260                          0.008014
A2M                           7.459105                          9.279760
A2ML1                         1.126008                          0.083225
A4GALT                        5.249234                          5.123996

        TCGA-D8-A73X-01A-11R-A32P-07  TCGA-AR-A0TP-01A-11R-A084-07  \
A1BG                          4.324578                          3.687565
A1CF                          0.021112                          0.020984
A2M                           7.526283                         12.770294
A2ML1                         0.274456                          0.220169
A4GALT                        5.465881                          3.083101

        TCGA-E2-A1IF-01A-11R-A144-07  TCGA-EW-A6SD-01A-12R-A33J-
07  ...  \
A1BG                          2.399717
4.233443  ...
A1CF                          0.025614
0.108061  ...
A2M                           8.201515
8.065138  ...
A2ML1                         0.152375
1.333526  ...
A4GALT                        5.371876
3.027163  ...

        TCGA-N5-A4RF-01A-11R-A28V-07  TCGA-N6-A4VF-01A-31R-A28V-07  \
A1BG                          4.434854                          3.630666
A1CF                          0.044659                          0.007317
A2M                           9.212672                         10.137570
A2ML1                         0.063074                          0.546730
A4GALT                        4.821015                          5.287393

        TCGA-N5-A4RN-01A-12R-A28V-07  TCGA-QM-A5NM-01A-11R-A28V-07  \
A1BG                          3.575404                          2.458971
A1CF                          0.015572                          0.035358
A2M                           8.439769                          5.508460
```

```
A2ML1                          0.744056                          2.409112
A4GALT                         4.690114                          4.438399

        TCGA-N5-A4RJ-01A-11R-A28V-07  TCGA-N5-A4RO-01A-11R-A28V-07  \
A1BG                           5.791871                          3.535440
A1CF                           1.949092                          0.078157
A2M                            6.958057                          7.928136
A2ML1                          0.686557                          0.088065
A4GALT                         4.914086                          3.492874

        TCGA-N5-A4RV-01A-21R-A28V-07  TCGA-N6-A4VD-01A-11R-A28V-07  \
A1BG                           6.075494                          4.383862
A1CF                           0.021624                          0.006336
A2M                            9.668470                          8.952802
A2ML1                          3.002216                          0.977974
A4GALT                         2.089483                          5.706015

        TCGA-N5-A4RT-01A-11R-A28V-07  TCGA-ND-A4WC-01A-21R-A28V-07
A1BG                           4.488649                          3.182423
A1CF                           0.019464                          0.091276
A2M                            6.137774                          7.317486
A2ML1                          0.568217                          0.012821
A4GALT                         4.227761                          5.491313

[5 rows x 1802 columns]
(15716, 1802)
<class 'pandas.core.frame.DataFrame'>
Index: 15716 entries, A1BG to ZZZ3
Columns: 1802 entries, TCGA-E9-A1NI-01A-11R-A14D-07 to TCGA-ND-A4WC-
01A-21R-A28V-07
dtypes: float64(1802)
memory usage: 216.2+ MB
None
       TCGA-E9-A1NI-01A-11R-A14D-07  TCGA-E2-A1LK-01A-21R-A14D-07  \
count                  15716.000000                  15716.000000
mean                       3.819608                      3.695116
std                        2.367493                      2.448152
min                        0.000000                      0.000000
25%                        1.936472                      1.630385
50%                        3.906571                      3.775523
75%                        5.435952                      5.426145
max                       12.964224                     14.202553

       TCGA-BH-A0B2-01A-11R-A10J-07  TCGA-E2-A107-01A-11R-A10J-07  \
count                  15716.000000                  15716.000000
mean                       3.990414                      3.820167
std                        2.285236                      2.453945
min                        0.000000                      0.000000
25%                        2.257207                      1.747294
50%                        4.136273                      3.930274
```

```
75%                          5.549962                          5.587596
max                         13.443264                         13.090924

       TCGA-LL-A5YN-01A-11R-A28M-07  TCGA-BH-A0DQ-01A-11R-A084-07  \
count                  15716.000000                  15716.000000
mean                       3.605881                      4.011386
std                        2.355658                      2.326681
min                        0.000000                      0.000000
25%                        1.711473                      2.235622
50%                        3.516790                      4.192228
75%                        5.211256                      5.601328
max                       14.282765                     12.850413

       TCGA-D8-A73X-01A-11R-A32P-07  TCGA-AR-A0TP-01A-11R-A084-07  \
count                  15716.000000                  15716.000000
mean                       3.989347                      3.887523
std                        2.340917                      2.395201
min                        0.000000                      0.000000
25%                        2.162130                      1.930483
50%                        4.200889                      4.070886
75%                        5.631869                      5.570968
max                       14.598147                     13.806524

       TCGA-E2-A1IF-01A-11R-A144-07  TCGA-EW-A6SD-01A-12R-A33J-07  ...
\
count                  15716.000000                  15716.000000  ...

mean                       3.749432                      3.807983  ...

std                        2.324721                      2.368889  ...

min                        0.000000                      0.000000  ...

25%                        1.925552                      1.848337  ...

50%                        3.785585                      3.938842  ...

75%                        5.353088                      5.461877  ...

max                       15.469026                     14.315320  ...


       TCGA-N5-A4RF-01A-11R-A28V-07  TCGA-N6-A4VF-01A-31R-A28V-07  \
count                  15716.000000                  15716.000000
mean                       3.550416                      3.983453
std                        2.392339                      2.259327
min                        0.000000                      0.000000
25%                        1.572896                      2.324922
50%                        3.485249                      4.075362
75%                        5.189607                      5.489849
```

```
max                          14.453351                          13.782161

       TCGA-N5-A4RN-01A-12R-A28V-07  TCGA-QM-A5NM-01A-11R-A28V-07  \
count                  15716.000000                  15716.000000
mean                       3.788677                      3.393325
std                        2.345979                      2.388829
min                        0.000000                      0.000000
25%                        1.918349                      1.421703
50%                        3.895144                      3.229124
75%                        5.440074                      4.995577
max                       13.547581                     14.285496

       TCGA-N5-A4RJ-01A-11R-A28V-07  TCGA-N5-A4RO-01A-11R-A28V-07  \
count                  15716.000000                  15716.000000
mean                       3.766908                      3.798477
std                        2.317204                      2.337397
min                        0.000000                      0.000000
25%                        1.963479                      1.893287
50%                        3.703291                      3.899546
75%                        5.343319                      5.440942
max                       13.783820                     13.375119

       TCGA-N5-A4RV-01A-21R-A28V-07  TCGA-N6-A4VD-01A-11R-A28V-07  \
count                  15716.000000                  15716.000000
mean                       3.440556                      4.015677
std                        2.446317                      2.240938
min                        0.000000                      0.000000
25%                        1.339680                      2.370121
50%                        3.272553                      4.110397
75%                        5.136387                      5.525538
max                       14.163920                     13.023853

       TCGA-N5-A4RT-01A-11R-A28V-07  TCGA-ND-A4WC-01A-21R-A28V-07
count                  15716.000000                  15716.000000
mean                       4.029607                      3.729629
std                        2.289467                      2.367788
min                        0.000000                      0.000000
25%                        2.320842                      1.807498
50%                        4.103340                      3.810268
75%                        5.616946                      5.354473
max                       13.812769                     13.594760

[8 rows x 1802 columns]
<class 'pandas.core.frame.DataFrame'>
Index: 1802 entries, TCGA-E9-A1NI-01A-11R-A14D-07 to TCGA-ND-A4WC-01A-
21R-A28V-07
Data columns (total 71 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   cancer_type                             1802 non-null   object
```

```
1   bcr_patient_barcode                     1730 non-null   object
2   bcr_patient_uuid                        1730 non-null   object
3   patient_id                              1730 non-null   object
4   gender                                  1730 non-null   object
5   race                                    1730 non-null   object
6   ethnicity                               1730 non-null   object
7   age_at_diagnosis                        694 non-null    float64
8   age_at_initial_pathologic_diagnosis     1036 non-null   object
9   birth_days_to                           1594 non-null   object
10  last_contact_days_to                    1594 non-null   object
11  death_days_to                           1594 non-null   object
12  clinical_stage                          1219 non-null   object
13  clinical_T                              1219 non-null   object
14  clinical_N                              1219 non-null   object
15  clinical_M                              1141 non-null   object
16  ajcc_pathologic_tumor_stage             1144 non-null   object
17  ajcc_staging_edition                    1142 non-null   object
18  ajcc_nodes_pathologic_pn                1142 non-null   object
19  ajcc_metastasis_pathologic_pm           1142 non-null   object
20  ajcc_tumor_pathologic_pt                1142 non-null   object
21  pathologic_stage                        366 non-null    object
22  ajcc_clinical_tumor_stage               225 non-null    object
23  tumor_status                            1650 non-null   object
24  tumor_type                              73 non-null     object
25  histologic_diagnosis                    1272 non-null   object
26  neoplasm_histologic_grade               153 non-null    object
27  tumor_grade                             536 non-null    object
28  nuclear_grade_III_IV                    79 non-null     object
29  residual_tumor                          908 non-null    object
30  residual_disease_largest_nodule         77 non-null     object
31  vital_status                            1730 non-null   object
32  days_to_death                           136 non-null    object
33  days_to_last_followup                   136 non-null    object
34  tissue_source_site                      1723 non-null   object
35  tissue_source_site.1                    0 non-null      float64
36  tumor_tissue_site                       1651 non-null   object
37  tumor_site                              70 non-null     object
38  anatomic_neoplasm_subdivision           588 non-null    object
39  anatomic_neoplasm_subdivision_other     232 non-null    object
40  submitted_tumor_site                    79 non-null     object
41  metastatic_tumor_site                   158 non-null    object
42  metastatic_tumor_site.1                 79 non-null     object
43  history_other_malignancy                1730 non-null   object
44  history_thyroid_disease                 79 non-null     object
45  history_thyroid_disease_other           79 non-null     object
46  history_reflux_disease_indicator        80 non-null     object
47  history_exposure_leukemogenic_agents    16 non-null     object
48  history_neoadjuvant_treatment           1730 non-null   object
49  radiation_treatment_adjuvant            1650 non-null   object
```

```
 50  treatment_outcome_first_course        1339 non-null   object
 51  pharmaceutical_tx_adjuvant            1053 non-null   object
 52  platelet_count                        217 non-null    object
 53  platelet_count_preresection           157 non-null    object
 54  platelet_norm_range_lower             77 non-null     object
 55  platelet_norm_range_upper             77 non-null     object
 56  icd_o_3_site                          1730 non-null   object
 57  icd_o_3_histology                     1730 non-null   object
 58  icd_10                                1730 non-null   object
 59  disease_code                          1730 non-null   object
 60  project_code                          1730 non-null   object
 61  ajcc_pathologic_t                     0 non-null      float64
 62  history_hormonal_contraceptives_use   211 non-null    object
 63  pregnancies_count_miscarriage         77 non-null     object
 64  jewish_religion_heritage_indicator    77 non-null     object
 65  tobacco_smoking_age_started           153 non-null    object
 66  family_history_cancer_type            76 non-null     object
 67  tumor_response                        77 non-null     object
 68  ecog_score                            896 non-null    object
 69  lymph_nodes_examined_count            991 non-null    object
 70  residual_tumor.1                      908 non-null    object
dtypes: float64(3), object(68)
memory usage: 1013.6+ KB
None
       age_at_diagnosis  tissue_source_site.1  ajcc_pathologic_t
count        694.000000                   0.0                0.0
mean          57.680115                   NaN                NaN
std           14.887310                   NaN                NaN
min           15.000000                   NaN                NaN
25%           48.000000                   NaN                NaN
50%           58.500000                   NaN                NaN
75%           69.000000                   NaN                NaN
max           90.000000                   NaN                NaN
      TCGA-E9-A1NI-01A-11R-A14D-07  TCGA-E2-A1LK-01A-21R-A14D-07  \
TERT                      0.121333                      0.881043

      TCGA-BH-A0B2-01A-11R-A10J-07  TCGA-E2-A107-01A-11R-A10J-07  \
TERT                      0.101666                      0.531254

      TCGA-LL-A5YN-01A-11R-A28M-07  TCGA-BH-A0DQ-01A-11R-A084-07  \
TERT                      0.741782                      0.323436

      TCGA-D8-A73X-01A-11R-A32P-07  TCGA-AR-A0TP-01A-11R-A084-07  \
TERT                      0.113267                      1.156543

      TCGA-E2-A1IF-01A-11R-A144-07  TCGA-EW-A6SD-01A-12R-A33J-07  ...
\
TERT                      0.200219                      0.352796  ...
```

```
      TCGA-N5-A4RF-01A-11R-A28V-07  TCGA-N6-A4VF-01A-31R-A28V-07  \
TERT                        0.1819                      0.195208

      TCGA-N5-A4RN-01A-12R-A28V-07  TCGA-QM-A5NM-01A-11R-A28V-07  \
TERT                      0.917643                       0.38555

      TCGA-N5-A4RJ-01A-11R-A28V-07  TCGA-N5-A4RO-01A-11R-A28V-07  \
TERT                      0.966262                      1.953097

      TCGA-N5-A4RV-01A-21R-A28V-07  TCGA-N6-A4VD-01A-11R-A28V-07  \
TERT                      0.252477                      1.002368

      TCGA-N5-A4RT-01A-11R-A28V-07  TCGA-ND-A4WC-01A-21R-A28V-07
TERT                      1.759906                      0.428633

[1 rows x 1802 columns]
       TCGA-E9-A1NI-01A-11R-A14D-07  TCGA-E2-A1LK-01A-21R-A14D-07  \
count                      1.000000                      1.000000
mean                       0.121333                      0.881043
std                             NaN                           NaN
min                        0.121333                      0.881043
25%                        0.121333                      0.881043
50%                        0.121333                      0.881043
75%                        0.121333                      0.881043
max                        0.121333                      0.881043

       TCGA-BH-A0B2-01A-11R-A10J-07  TCGA-E2-A107-01A-11R-A10J-07  \
count                      1.000000                      1.000000
mean                       0.101666                      0.531254
std                             NaN                           NaN
min                        0.101666                      0.531254
25%                        0.101666                      0.531254
50%                        0.101666                      0.531254
75%                        0.101666                      0.531254
max                        0.101666                      0.531254

       TCGA-LL-A5YN-01A-11R-A28M-07  TCGA-BH-A0DQ-01A-11R-A084-07  \
count                      1.000000                      1.000000
mean                       0.741782                      0.323436
std                             NaN                           NaN
min                        0.741782                      0.323436
25%                        0.741782                      0.323436
50%                        0.741782                      0.323436
75%                        0.741782                      0.323436
max                        0.741782                      0.323436

       TCGA-D8-A73X-01A-11R-A32P-07  TCGA-AR-A0TP-01A-11R-A084-07  \
count                      1.000000                      1.000000
mean                       0.113267                      1.156543
std                             NaN                           NaN
```

```
min                          0.113267                          1.156543
25%                          0.113267                          1.156543
50%                          0.113267                          1.156543
75%                          0.113267                          1.156543
max                          0.113267                          1.156543

       TCGA-E2-A1IF-01A-11R-A144-07  TCGA-EW-A6SD-01A-12R-A33J-07  ...
\
count                        1.000000                          1.000000  ...

mean                         0.200219                          0.352796  ...

std                               NaN                               NaN  ...

min                          0.200219                          0.352796  ...

25%                          0.200219                          0.352796  ...

50%                          0.200219                          0.352796  ...

75%                          0.200219                          0.352796  ...

max                          0.200219                          0.352796  ...


       TCGA-N5-A4RF-01A-11R-A28V-07  TCGA-N6-A4VF-01A-31R-A28V-07  \
count                        1.0000                          1.000000
mean                         0.1819                          0.195208
std                             NaN                               NaN
min                          0.1819                          0.195208
25%                          0.1819                          0.195208
50%                          0.1819                          0.195208
75%                          0.1819                          0.195208
max                          0.1819                          0.195208

       TCGA-N5-A4RN-01A-12R-A28V-07  TCGA-QM-A5NM-01A-11R-A28V-07  \
count                        1.000000                           1.00000
mean                         0.917643                           0.38555
std                               NaN                               NaN
min                          0.917643                           0.38555
25%                          0.917643                           0.38555
50%                          0.917643                           0.38555
75%                          0.917643                           0.38555
max                          0.917643                           0.38555

       TCGA-N5-A4RJ-01A-11R-A28V-07  TCGA-N5-A4RO-01A-11R-A28V-07  \
count                        1.000000                          1.000000
mean                         0.966262                          1.953097
std                               NaN                               NaN
min                          0.966262                          1.953097
```

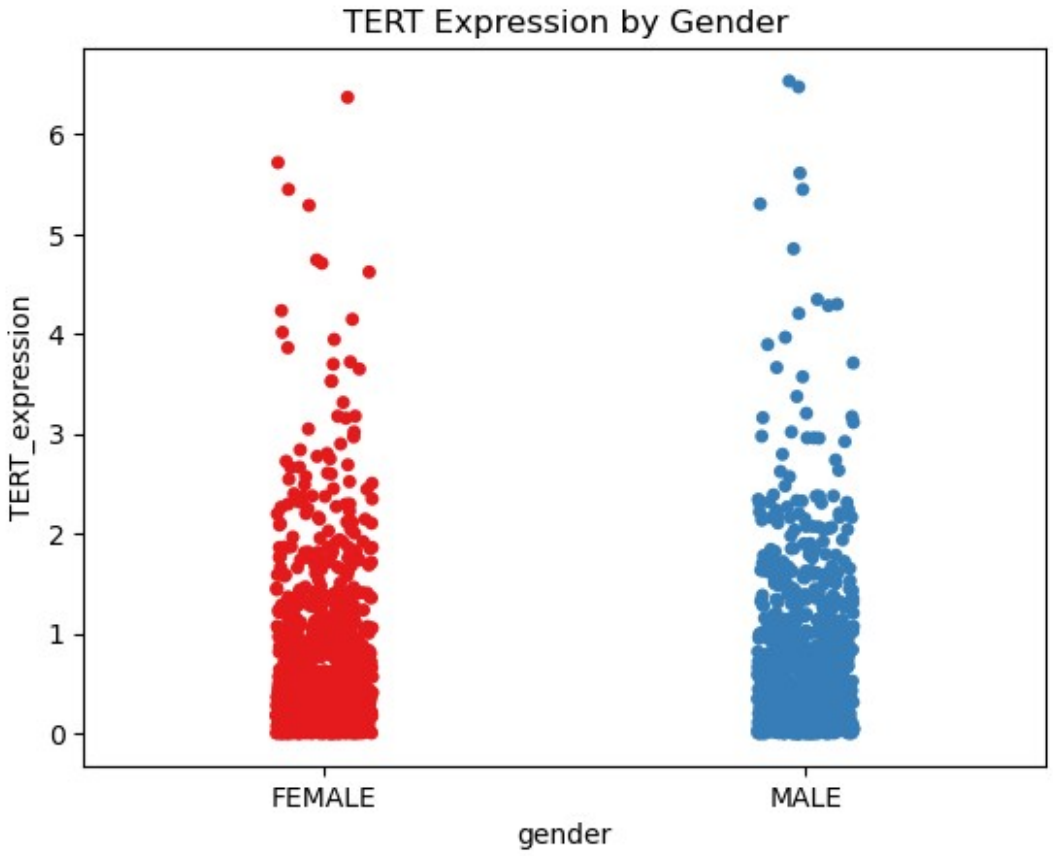
```
25%                            0.966262                            1.953097
50%                            0.966262                            1.953097
75%                            0.966262                            1.953097
max                            0.966262                            1.953097

       TCGA-N5-A4RV-01A-21R-A28V-07  TCGA-N6-A4VD-01A-11R-A28V-07  \
count                      1.000000                      1.000000
mean                       0.252477                      1.002368
std                             NaN                           NaN
min                        0.252477                      1.002368
25%                        0.252477                      1.002368
50%                        0.252477                      1.002368
75%                        0.252477                      1.002368
max                        0.252477                      1.002368

       TCGA-N5-A4RT-01A-11R-A28V-07  TCGA-ND-A4WC-01A-21R-A28V-07
count                      1.000000                      1.000000
mean                       1.759906                      0.428633
std                             NaN                           NaN
min                        1.759906                      0.428633
25%                        1.759906                      0.428633
50%                        1.759906                      0.428633
75%                        1.759906                      0.428633
max                        1.759906                      0.428633

[8 rows x 1802 columns]
TERT    0.740426
dtype: float64
TERT    0.698498
dtype: float64
TERT    0.394489
dtype: float64
```

```python
### Investigate potential correlation between TERT and gender
#Import necessary modules
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix

# Clean the training data of NAN values
non_NAN = metadata_df['gender'].dropna().index
X1 = gene_data.loc['TERT', non_NAN]
y1 = metadata_df.loc[non_NAN, 'gender']

#Convert gender to binary to quantify gender feature
y_label = [{"MALE": 0, "FEMALE": 1}[i] for i in y1]

# Combine TERT expression and gender into a DataFrame
plot_df = pd.DataFrame({
    "TERT_expression": X1.values,
```

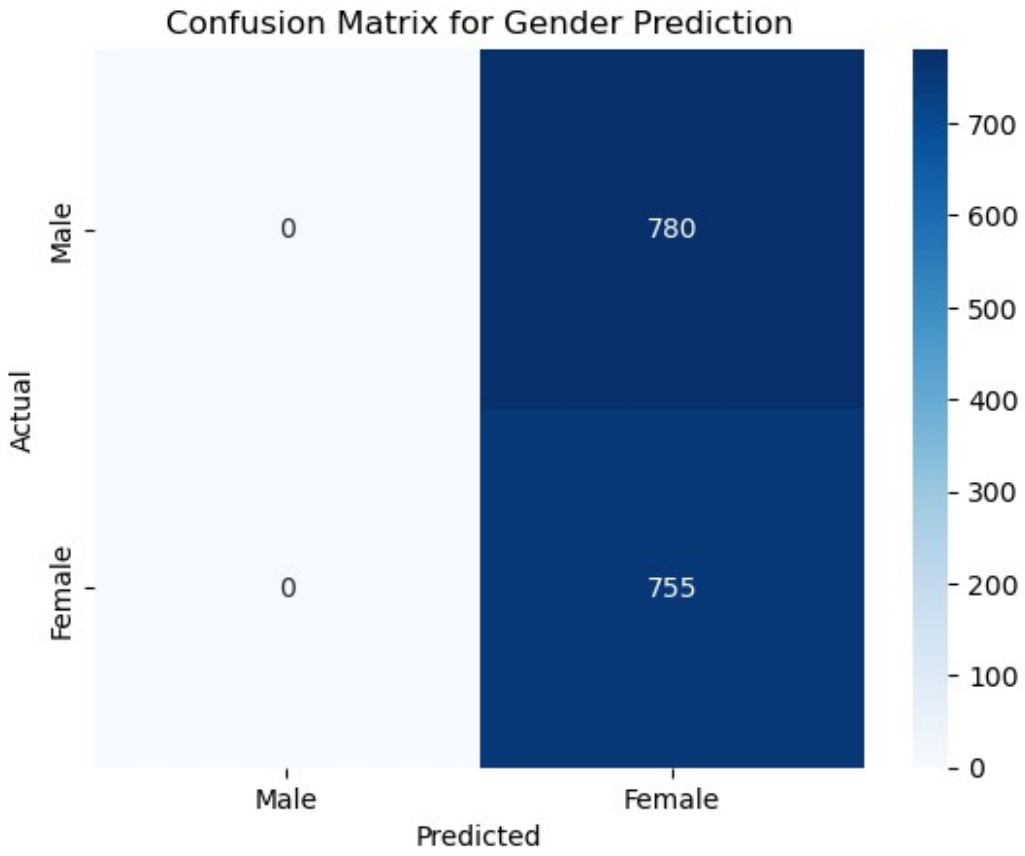
```
    "gender": y1.values
}, index=y1.index)

# Plot TERT against gender with stripplot
sns.stripplot(data=plot_df,
              x="gender",
              y="TERT_expression",
              palette="Set1")
plt.title("TERT Expression by Gender")
plt.show()

#Load testing data set
test_data = pd.read_csv(
'/Users/Melanie/Downloads/Comp_BME/Module3_Cancer/TEST_SET_GSE62944_su
bsample_log2TPM.csv', index_col=0, header=0)
test_meta = pd.read_csv(
'/Users/Melanie/Downloads/Comp_BME/Module3_Cancer/TEST_SET_GSE62944_me
tadata.csv', index_col=0, header=0)

#Clean the testing data set
test_non_NAN = test_meta['gender'].dropna().index
test_X1 = test_data.loc['TERT', test_non_NAN]
test_y1 = test_meta.loc[test_non_NAN, 'gender']

#Train Logistic regression then test with testing data set
X1_log = X1.values.reshape(-1,1)
model = LogisticRegression().fit(X1_log, y_label)
test_X1_log = test_X1.values.reshape(-1,1)
pred_gender = model.predict(test_X1_log)

# Encode test labels to binary
test_y_label = [{"MALE": 0, "FEMALE": 1}[i] for i in test_y1]
test_y1_binary = np.array(test_y_label)

# Compute accuracy and confusion matrix
print(f"Model accuracy: {model.score(test_X1_log, test_y1_binary)}")
cm = confusion_matrix(test_y1_binary, pred_gender)

# Plot heatmap of confusion matrix
sns.heatmap(cm, annot=True, fmt='d', cmap="Blues",
            xticklabels=["Male", "Female"],
            yticklabels=["Male", "Female"])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix for Gender Prediction")
plt.show()
```

```
/var/folders/9_/7k0ty_k13q13lv0m0h2syp3w0000gn/T/
ipykernel_15580/2441500197.py:22: FutureWarning:
```

```
Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.stripplot(data=plot_df,
```




```
Model accuracy: 0.49185667752442996
```

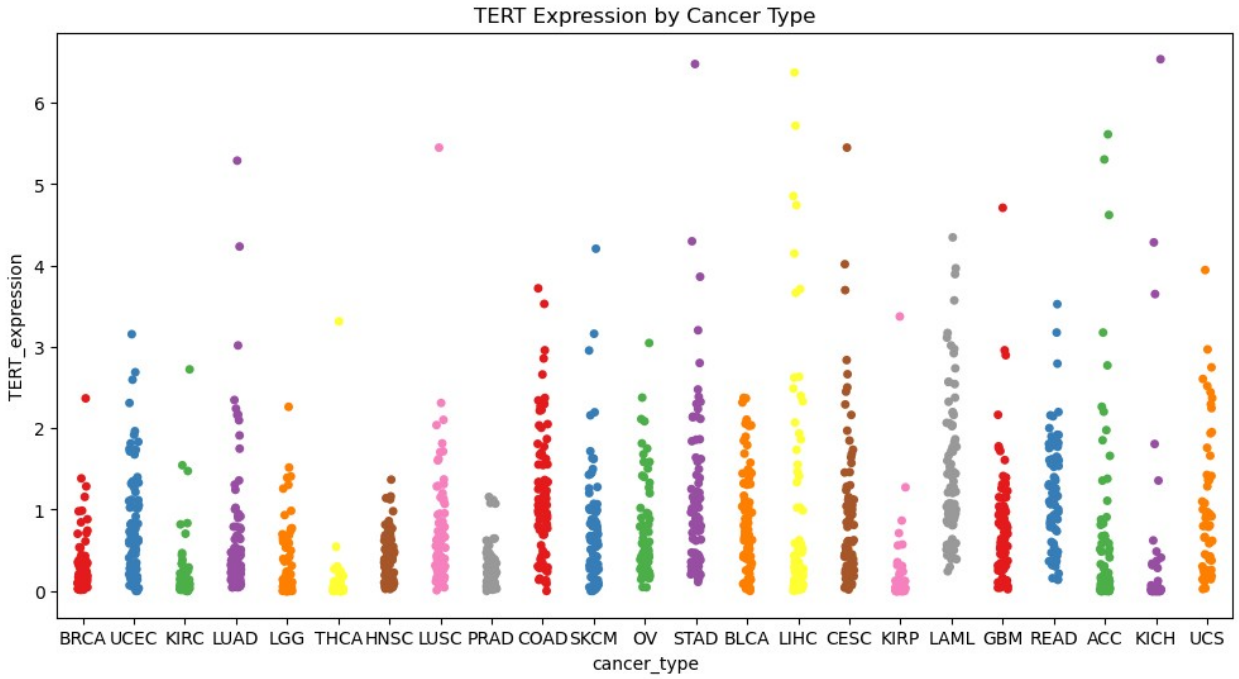



```
### Investigate potential correlation between TERT and cancer types
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import classification_report

# Clean the data of NAN values
non_NAN = metadata_df['cancer_type'].dropna().index
X2 = gene_data.loc['TERT',non_NAN]
y2 = metadata_df.loc[non_NAN, 'cancer_type']

#Clean the testing data set
test_non_NAN = test_meta['cancer_type'].dropna().index
test_X2 = test_data.loc['TERT', test_non_NAN]
test_y2 = test_meta.loc[test_non_NAN, 'cancer_type']

# Convert cancer types into numeric labels for both traning and
testing sets
le_all = LabelEncoder()
le_all.fit(pd.concat([y2, test_y2]))

# Transform train and test labels consistently
y_quant2 = le_all.transform(y2)
test_y_quant2 = le_all.transform(test_y2)
```

```
# Combine TERT expression and cancer type into a DataFrame
plot_df2 = pd.DataFrame({
    "TERT_expression": X2.values,
    "cancer_type": y2.values
}, index=y2.index)

# Plot TERT against cancer type with stripplot
plt.figure(figsize=(12,6))
sns.stripplot(data=plot_df2,
              x="cancer_type",
              y="TERT_expression",
              palette="Set1")
plt.title("TERT Expression by Cancer Type")
plt.show()

#Train Logistic regression then test with testing data set
X2_log = X2.values.reshape(-1,1)
model = LogisticRegression(multi_class='multinomial', solver='lbfgs',
max_iter=1000)
model.fit(X2_log, y_quant2)
test_X2_log = test_X2.values.reshape(-1,1)
pred_cancer = model.predict(test_X2_log)

# Compute classification report and confusion matrix
unique_classes = np.unique(test_y_quant2) # Get the unique classes
present in the test set
print(f"Classification report: {classification_report(test_y_quant2,
pred_cancer,

labels=unique_classes,

target_names=le_all.classes_[unique_classes])}")
cm = confusion_matrix(test_y_quant2, pred_cancer)

#Plot heatmap of confusion matrix
sns.heatmap(cm, annot=True, fmt='d', cmap="Blues",
            xticklabels=le_all.classes_,
            yticklabels=le_all.classes_)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix for Cancer Type Prediction")
plt.show()
```

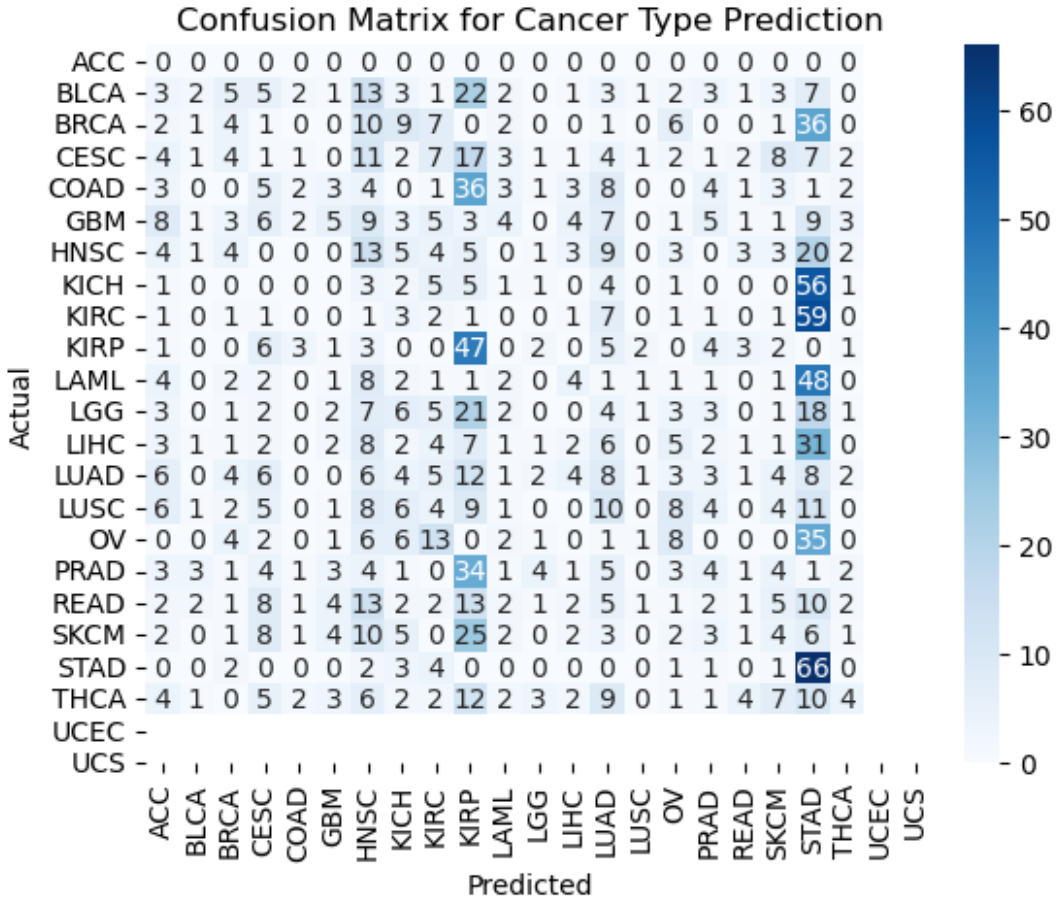
```
/var/folders/9_/7k0ty_k13q13lv0m0h2syp3w0000gn/T/
ipykernel_15580/1049369323.py:31: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.
```

```
  sns.stripplot(data=plot_df2,
```




```
/Users/Melanie/Documents/anaconda3/lib/python3.13/site-packages/
sklearn/linear_model/_logistic.py:1247: FutureWarning: 'multi_class'
was deprecated in version 1.5 and will be removed in 1.7. From then
on, it will always use 'multinomial'. Leave it to its default value to
avoid this warning.
  warnings.warn(

Classification report:                precision    recall  f1-score
support

        BLCA       0.14      0.03      0.04        80
        BRCA       0.10      0.05      0.07        80
        CESC       0.01      0.01      0.01        80
        COAD       0.13      0.03      0.04        80
         GBM       0.16      0.06      0.09        80
        HNSC       0.09      0.16      0.12        80
        KIRC       0.03      0.03      0.03        80
        KIRP       0.03      0.03      0.03        80
        LAML       0.17      0.59      0.27        80
         LGG       0.06      0.03      0.04        80
        LIHC       0.00      0.00      0.00        80
        LUAD       0.07      0.03      0.04        80
        LUSC       0.08      0.10      0.09        80
          OV       0.00      0.00      0.00        80
        PRAD       0.15      0.10      0.12        80
```

```
        READ       0.10      0.05      0.07        80
        SKCM       0.05      0.01      0.02        80
        STAD       0.07      0.05      0.06        80
        THCA       0.15      0.82      0.25        80
        UCEC       0.17      0.05      0.08        80

   micro avg       0.11      0.11      0.11      1600
   macro avg       0.09      0.11      0.07      1600
weighted avg       0.09      0.11      0.07      1600
```



Verify and validate your analysis:

The methods we used to validate our data were model accuracy scores, reports, and confusion matrices.

The model accuracy score uses the determined optimized loss function and calculates out of how many predictions were correct when compared with the actual data. For the TERT vs. Gender classification model, the model predicted around half correct, while the other half was incorect. Since the data was binary, the model was as good as randomly guessing between the two given genders.

The classification report uses the determined groupings to determine how many predictions were correct, and then using those figures to calculate the precision, recall, f1-score, and support. For the TERT vs. Cancer Type classification model, both the precision and recall rates were very low, meaning that the model was not accurate in predicting cancer type from TERT.

Confusion matrices were used for both the classification models as confusion matrices provide a practical way to study how well the models used TERT to classify the respective feature. It shows, very similarly to the aforementioned validation methods, how many predictions were mapped to the correct group, and how many predictions were falsely labeled as another. The model accuracy score for TERT vs. Gender is supported by its confusion matrix as the model correctly predicted half of the data as females, though falsely labeled the other half as male. The confusion matrix for the cancer types is slightly more complicated due to the many types of cancers in the dataset, and this complexity is reflected in the model accuracy itself. The number of false predictions are scattered across the confusion matrix, with STAD cancer having the highest number of true predictions. Oddly enough, there are also a lot of false predictions of STAD that were either KIRC or KICH. Overall, the matrix shows that the classification model's accuracy is poor as well.

Current literature has found that there are differing rates of TERT promoter mutations across different sexes shown in certain cancers, meaning that there could potentially be different amounts of TERT expression based on gender and cancer type. However, these trends are so far only found in specific studies such as gender and melanoma, gender and thyroid cancer, and gender and hepatocellular carcinoma (also known as liver cancer) (El Zarif et al.). These studies are much narrower than our method, which compared TERT expression across nearly 20 types of cancers independent of gender, then TERT expression across gender independent of cancer type. The few classifications that were predicted correctly may have coincidentally aligned with the specific studies that have linked TERT expression by gender to certain cancers, or may have predicted them correctly by chance. Additionally, TERT overexpression in cancer has only been shown to be found in 4% of cancers (Colebatch et al.), which could explain why using solely TERT as a biomarker to classify cancers in our model did not provide adequate results.

- Talal El Zarif, Marc Machaalani, Rashad Nawfal, Amin H Nassar, Wanling Xie, Toni K Choueiri, Mark Pomerantz, TERT Promoter Mutations Frequency Across Race, Sex, and Cancer Type, The Oncologist, Volume 29, Issue 1, January 2024, Pages 8–14, https://doi.org/10.1093/oncolo/oyad208
- Colebatch AJ, Dobrovic A, Cooper WATERT gene: its function and dysregulation in cancerJournal of Clinical Pathology 2019;72:281-284.

Conclusions and Ethical Implications:

CONCLUSION

Across the TCGA Pan-Cancer cohort, TERT expression alone did not yield predictive power for either sex classification or cancer-type classification. The TERT→sex model performed at ~50% accuracy (essentially random guessing for a binary task) and the TERT→cancer-type model showed uniformly poor precision and recall across >15 cancer types.

This aligns with prior research showing that while TERT can be profoundly important in specific tumor contexts, it is not universally dysregulated across cancer. Colebatch et al. report that clearly oncogenic TERT promoter driven-overexpression is only present in a minority of human

tumors (≈4%) (Colebatch et al., 2019). The pan-cancer null result we saw is actually consistent: most cancers do not depend on telomerase activation via promoter mutation for malignant fitness. In some cancer classes (e.g. ALT-driven sarcomas) the tumor does not rely on telomerase at all.

ETHICAL IMPLICATIONS

No sex-stratified inference is justified. Our model showed no generalizable linkage between sex assignment and TERT expression with averaged pan-cancer. This matters ethically because there are published studies showing sex-differences within specific cancers (e.g. hepatocellular carcinoma (Nault et al., 2019), melanoma (Hodis et al., 2022)), but those findings cannot be naïvely generalized.

Single-biomarker AI should not be used for clinical triage. The idea that “high TERT” tells you “which cancer this is” or “which sex the patient is” is scientifically unsound. An AI built on only TERT would risk misclassification harm. Regulatory AI ethics frameworks (e.g. in oncology FDA guidance) explicitly warn against single-feature clinical models.

Context matters in genomics. TERT must not be interpreted without the other pathway context (p53, ATRX/DAXX, RAS/RAF pathway status, epigenome). Prior studies that did show strong TERT effects showed them in mechanistically coherent domains, not as standalone signals.

Limitations and Future Work:

LIMITATIONS

TCGA sex field is biological sex — TCGA does not have gender-identity, hormonal therapy exposure, menopausal state or gonadal hormonal axis variables (all of which strongly modulate telomerase activity). Telomerase activity is known to be estrogen-responsive in normal physiology and carcinogenesis (Bayne & Liu, Nature, 2020). Without hormone axis variables, our model cannot capture the whole picture.

Cancer-type classification is extremely imbalanced in TCGA. Some cancer classes have N>1000 (e.g. BRCA) while rare cancers have N<100. Classifiers break under such regimes. Only a single molecular feature (TERT) was used — highly unlikely to be a sufficient biomarker. TCGA is not a random population sample — it is a convenience cohort and has known race and stage biases. Cancer-type classification is a multi-class task with highly imbalanced support in TCGA — most cancers have very small N. TERT expression is not normalized to tumor purity. Without purity correction (ABSOLUTE / ESTIMATE / CIBERSORT), you cannot tell transcription from cancer cells vs tumor microenvironment stromal cells. Additionally, there is no molecular subtype stratification. For example, BRCA has LumA / LumB / HER2 / basal-like — those subtypes have different telomerase pathway wiring. If you collapse across subtype, you destroy signal before modeling even begins.

TCGA is pre-immunotherapy era (samples collected ~2005–2014). Modern tumors under immune checkpoint blockade adapt telomere biology differently (Bussani et al., Cancer Cell, 2021), so the generalization to modern clinics may be limited.

FUTURE WORK

Future iterations should not treat TERT as a stand-alone scalar feature. Instead, next work must move toward systems-level telomere biology modeling.

First, models should be cancer-type specific, not pan-cancer. Nearly every major telomerase study that has found a sex difference in TERT (melanoma, thyroid, HCC) showed those differences within a cancer type — not averaged across the entire human oncologic space. This mirrors what the Pan-Cancer Analysis of Whole Genomes consortium demonstrated: TERT-activating alterations are strongly tumor-type specific (PCAWG Consortium, Nature, 2020).

Second, future modeling should include other molecular axis variables required for telomerase biology: TERT promoter mutation calls (which are more stable than RNA expression; Huang et al., Nat Genet, 2023), ATRX/DAXX status, TP53 disruption, ALT phenotype estimates, and bulk telomere length metrics (Barthel et al., Nat Genet, 2017). Telomerase activation is not a single-gene RNA phenomenon — it is a circuit-level state.

Third, confounders must be explicitly modeled — especially age, hormonal exposure, and tumor purity. Telomerase activity is estrogen-modulated and age-modulated (Bayne & Liu, Nature, 2020), and TCGA sex is a coarse binary field that does not represent nuanced endocrine biology. Without controlling for these axes, any future sex-based inference will remain uninterpretable.

Finally, methodological advances should be used. Graph neural networks, Bayesian networks, or Cox-based deep survival models (DeepSurv / DeepHit) would allow modeling of telomere pathway state as a structured system, not a single marker. There is growing evidence that the future of precision oncology is multi-omic telomere-state modeling, not single gene biomarkers.

Barthel FP, Wei W, Tang M, et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. Nature Genetics. 2017;49(3):349–357. doi:10.1038/ng.3781

Bayne S, Liu JP. Telomeres and telomerase: from discovery to clinical cancer trials. Nature Review Molecular Cell Biology. 2020;21:407–428. doi:10.1038/s41580-020-0233-7

Huang FW, Hodis E, Xu MJ, Strickland MR, et al. Functional landscapes of TERT promoter mutations in cancer. Nature Genetics. 2023;55:1408–1417. doi:10.1038/s41588-023-01417-w

Nault JC, Martin Y, Caruso S, Hirsch TZ, et al. Clinical Impact of TERT Promoter Mutations in Hepatocellular Carcinoma. New England Journal of Medicine. 2019;380:2018-2020. doi:10.1056/NEJMc1904522

PCAWG Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93. doi:10.1038/s41586-020-1969-6

Zhou X, Zhang S. Telomerase reverse transcriptase in cancer molecular mechanisms and potential therapeutic target. Nature Reviews Molecular Cell Biology. 2022;23:318–337. doi:10.1038/s41580-021-00452-y

NOTES FROM YOUR TEAM:

- 10/21: Learned about optimization code and applications from guest lecturer Lavie
- 10/23: Met partner, learned about cancer, its hallmarks, and available datasets, decided on what to investigate and what future steps to take by checkin

- 10/25: First checkin due, finished background info, dataset info, and confirming that relevant metadata and genes are available in datasets
- 10/28: Learned about supervised machine learning
- 10/30: Learned about unsupervised machine learning
- 10/31-11/4: Implemented code to investigate question, ran into errors leading to changes in methods (regression to two separate classifications)
- 11/6: Learned about validation methods for supervised and unsupervised machine learning
- 11/7-8: Implemented a testing data set, improved visuals of data plots to make easier to understand, implemented valiation methods with accuracy scores and confusion matrices, found external validation
- 11/11: Learned about regularization
- 11/13: Added conclusions, ethical implications, and future work

QUESTIONS FOR YOUR TA:

N/A