

config.R

abdulbasir

2025-12-09

```
# configuration file for food & drug administration oncology approval analysis
#
# contains all constants, file paths, and parameters used throughout
# the analysis pipeline

# setting base directory paths
BASE_DIR = dirname(getwd())
DATA_DIR = file.path(BASE_DIR, "data")
RESULTS_DIR = file.path(BASE_DIR, "results")
FIGURES_DIR = file.path(BASE_DIR, "visualizations")

# creating output directories
dir.create(RESULTS_DIR, showWarnings = FALSE, recursive = TRUE)
dir.create(FIGURES_DIR, showWarnings = FALSE, recursive = TRUE)

# defining data file paths
RAW_DATA_FILE = file.path(
  DATA_DIR,
  "final_for_posting_compilation_of_cder_nme_and_new_biologic_approvals_1985-2024.csv"
)
CLEAN_DATA_FILE = file.path(RESULTS_DIR, "fda_analysis_clean.csv")

# oncology classification keywords
ONCOLOGY_KEYWORDS = c(
  "cancer", "tumor", "tumour", "carcinoma", "leukemia", "leukaemia",
  "lymphoma", "melanoma", "myeloma", "sarcoma", "oncology", "oncologic",
  "neoplasm", "neoplastic", "metastatic", "metastases", "malignant",
  "malignancy", "glioblastoma", "glioma", "blastoma", "adenocarcinoma",
  "squamous cell", "basal cell", "hodgkin", "myeloid", "lymphocytic",
  "mesothelioma", "astrocytoma", "neuroblastoma", "retinoblastoma",
  "hepatocellular", "cholangiocarcinoma", "renal cell", "urothelial",
  "multiple myeloma", "chronic myelogenous", "acute myeloid",
  "acute lymphoblastic", "non-small cell", "small cell",
  "thyroid cancer", "prostate cancer", "breast cancer",
  "ovarian cancer", "pancreatic cancer", "lung cancer",
  "colorectal cancer", "gastric cancer", "esophageal cancer",
  "chemotherapy", "antineoplastic", "cytotoxic"
)

# regulatory era cutoffs based on major fda legislation
ERA_CUTOFFS = list(
  "Pre-PDUFA" = c(1985, 1992), # prescription drug user fee act
```

```

"Early-PDUFA" = c(1993, 2002), # which is funding drug review via fees
"Mid-PDUFA" = c(2003, 2012),
"Post-FDASIA" = c(2013, 2024) # FDA safety and innovation act of 2012
) # law that reauthorized PDUFA & expanded FDA

# statistical parameters
ALPHA = 0.05
N_PERMUTATIONS = 10000
RANDOM_SEED = 123

# visualization parameters
DPI = 300
FIGURE_WIDTH = 12
FIGURE_HEIGHT = 8

# data quality thresholds
MAX REVIEW TIME DAYS = 3650 # 10 years
MIN REVIEW TIME DAYS = 0
INDEPENDENCE_THRESHOLD = 0.05 # 5 percent of sample per company
MIN CELL COUNT = 5 # minimum observations per cell for anova
MIN SAMPLE SIZE = 20 # minimum total n for era-stratified anova

# required columns for validation
REQUIRED_COLUMNS = c(
  "FDA Receipt Date",
  "FDA Approval Date",
  "Review Designation",
  "Abbreviated Indication(s)",
  "Orphan Drug Designation",
  "Accelerated Approval",
  "Breakthrough Therapy Designation",
  "Fast Track Designation",
  "Qualified Infectious Disease Product",
  "Approval Year"
)

```