Template for web interface for CANTRANce screening model

**Model type:** Screening

**Model title:** CANTRANce: Screening tests that Alter Stage Distributions at Incidence

Text to be added to the top of the page: The fundamental assumption underlying cancer screening is that early detection of a tumor will lead to correspondingly improved prognosis and disease-specific survival. A key consequence of early detection is the reduction in the incidence of advanced cancers associated with the screening test. Using the inputs specified below, CANTRANce will link comparative effectiveness study data on the reduction in the incidence of advanced tumors in the presence of a screening test into changes in disease-specific mortality and years of life saved under screening.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

FOR DEVELOPERS: This document is formatted the following way:

# SECTION HEADING

## Subsection heading

### Sub-subsection heading

#### ***Variable label***

{? Help text describing the type of information to be described in this variable or set of variables.}

**[Variable type]**

{Entry instructions: Description to help the user understand the specific values and formats required for each input.}

**${variable\_name}**

* Information about the variable, depending on its type. For freeform strings/tables, there are sections describing specifics for the web interface (Cornerstone) and for the downloadable interface (FHCRC) separately. General formatting information will be checked in the test\_inputs.r script no matter which type of interface is used, but if it’s possible to perform some of the checks in the interface itself, that would be fantastic.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# GENERAL INSTRUCTIONS

CANTRANce asks for inputs using either drop-down menus or entry boxes. Entry boxes are of three types:

1. **Numeric**. Enter a single number: e.g., 50

2. **String**. Enter text without using quotation marks: e.g., rx

3. **Table**. Enter rows of text and/or numbers, without using quotation marks, using commas (,) to separate the columns and new lines (\n) to separate rows. The first row should specify column names. This formatting can be achieved by constructing the table in Excel, saving the file as CSV (Comma Delimited), and then viewing the resulting .csv file in a simple text editor such as Notepad. Please view the defaults for examples.

You may choose to either run the model on the default dataset or use your own study data. The default data are simulated to represent FIXME

If you choose to use your own study data, it will be stored on a server along with the results of the models you run. Both the data and the results will be publicly accessible via a custom URL. If you require additional data privacy protection, please contact Jeanette Birnbaum about alternative modeling options.

All the default parameter inputs refer to the default dataset. Throughout the modeling process, time of study start is considered analogous to the time of randomization to trial arms in which individuals either receive preventive intervention (intervention group) or do not receive preventive intervention (control group).

# SIMULATION FEATURES

#### ***Number of simulations***

{? Each simulation represents one study linking the screening test to disease stage and cause-specific mortality, and multiple simulations capture the uncertainty in mortality estimation due to modeling assumptions. More simulations will improve uncertainty estimation but require a longer run-time. You may specify between 50 and 1,000 simulations. For experimental or practice runs, 100 simulations is recommended.}

**[Numeric]**

**${nsim}**

* Numeric
* Default=100
* Range=[50, 1000]

#### ***Years at which to evaluate survival***

{? CANTRANce will report k-year survival for the list of k’s provided here. The longest time period specified in this input will be used as the length of follow-up, or the number of years after study start for which the model will track mortality. For example, the default value will report 5-year, 10-year, and 50-year survival. Individuals in the study will be followed for 50 years. Lifetime follow-up can be approximated by specifying a maximum k by which all individuals should have died, e.g. 110-(average age).}

**[Table]**

{Entry instructions: Enter up to 6 numeric values between 0 and 100, separated by commas.}

**${times}**

* Freeform string
* Default=’5,10,50’
* Largest value must be between 1 and 100, inclusive

#### ***Population size***

{? The population size corresponds to the number of cases within each simulation. The maximum population size is 10,000.}

**[Numeric]**

**${pop\_size}**

* Numeric
* Default=500
* Range=[10,10000]

Set the study year as the time of randomization to trial arms. If individuals were enrolled over a period of multiple years, choose either the middle year or the year in which the majority of patients were enrolled

# COMPARATIVE EFFECTIVENESS DATA

#### ***Method for inputting data***

{? Select a method for specifying the study population. If individual-level data is available with all covariates of interest, choose *individual-level data*. Otherwise, you may choose *covariate proportions* and specify marginal distributions of covariates of interest. CANTRANce will simulate a study population accordingly.}

**[String]**

**${input\_method}**

* Choice string (drop-down menu)
* Allowable values:
  + individual\_data (label: Individual-level data)
  + covariate\_proportions (label: Covariate proportions)
* Default value:
  + covariate\_proportions

IF INDIVIDUAL DATA: fill out **Individual-level data** section.

IF COVARIATE PROPORTIONS: skip to **Covariate proportions** section.

## Individual-level data

#### ***Study data source***

{? Select an existing data source and skip to **Population Characteristics**, or select *user-provided individual-level data* and enter study data below.}

**[String]**

**${data\_source}**

* Choice string (drop-down menu)
* Allowable values:
  + default\_pc (label: Simulated prostate cancer data (default))
  + user\_data (label: User-provided individual-level data)
* Default value:
  + default\_pc
* *This option is not included explicitly in the R code*, but rather is a signal for the interface to either load default data or use user inputs. Default data is provided in the file named FIXME for the FIXME example for screening.

IF SIMULATED PROSTATE CANCER DATA: skip to **Sampling method.**

IF USER-PROVIDED INDIVIDUAL-LEVEL DATA: fill out **Study data.**

#### ***Study data***

{? The study data should be individual-level data describing patient characteristics that affect their risk of disease-specific death. CANTRANce assumes stage at clinical incidence is known for each individual. Stage at screened incidence will be assigned during the modeling process.}

**[Table]**

{Entry instructions: Rows should indicate individual cases and columns should specify relevant covariates. The following columns are required:

* *male* – sex of individual, with entries as 0 for female and 1 for male
* *age* OR *agegroup* – age of individual at study start
  + If *age* is specified:
    - Use single integers, e.g. 64
  + If *agegroup* is specified:
    - Use two integers separated by a dash, e.g. 60-64
    - The model will assign a single-year age to each individual according to a uniform distribution within the age range
* *stage* – disease stage at clinical incidence

The following information must be included either as a column in the input study data or as a separate input below:

* *study\_year* – year in which age was recorded

Additional columns may be included for covariates that have an effect on time from clinical incidence to disease-specific death.}

**${userdat\_file}**

* Default: FIXME
* For web interface:
  + Is it possible to have the user navigate to a file location and import the file path into the code? If not, have user copy and paste data into a text box
* For downloadable version:
  + File path (have user navigate to file location)
* Check for the following formatting in the resultant data table or file:
  + Column named *male*
    - Exists
    - Has values 0 AND/OR 1
  + Column for age
    - One or more of variables named *age* and *agegroup* exists
    - If *age* exists, entries should be formatted as integers
    - If *agegroup* exists, entries should be formatted as two integers separated by a dash
  + Column named *stage*
    - Exists

Sampling method

{? The sampling method determines how the simulated populations are sampled from the input data. Choose *Simple bootstrap* if you wish to preserve the covariate distribution present in the input data. The data can alternatively be resampled to represent a population with a different distribution of specified covariate strata; choose *Weighted bootstrap* and specify the desired proportion of individuals in of each stratum in the table below. The sum of the proportions must equal 1.

Note: To use the input database without modification, choose *Simple bootstrap* for the sampling method and make sure the **Population size** parameter above reflects the number of cases in each trial arm in the input database.}

**[String]**

**${create\_pop\_method}**

* Choice string (drop-down menu)
  + Allowable values:
    - simple\_bootstrap (label: Simple bootstrap)
    - weighted\_bootstrap (label: Weighted bootstrap)
  + Default value:
    - simple\_bootstrap

IF SIMPLE BOOTSTRAP: skip to **Stage shift.**

IF WEIGHTED BOOTSTRAP: fill out **Covariate distributions for weighted bootstrap sampling.**

#### Covariate distributions for weighted bootstrap sampling

**[Table]**

{Entry instructions: Specify a table of the distribution of covariate strata. Column names should correspond to the names of covariates in the input data, and rows should specify strata defined by combinations of those covariate values. The last column should be called *prop* with entries specifying the proportion of individuals in each stratum. The sum of the *prop* column must equal 1. For example, the default values correspond to the following table: <<Show HTML example of the default>>}

**${weighted\_bootstrap\_table}**

* For web interface:
  + Freeform string
    - Default=FIXME
* For downloadable version:
  + Allow user to choose desired covariates from among those in the input data that meet the following criteria:
    - Variable has 6 or fewer unique values
  + Populate a table with all unique combinations of values for the selected covariates and an additional (empty) column named *prop*.
    - For example, if user chose to include *male* (has unique values 0 and 1) and *history* (also has unique values 0 and 1), the table prompt would be:

|  |  |  |
| --- | --- | --- |
| Male | history | prop |
| 0 | 0 |  |
| 0 | 1 |  |
| 1 | 0 |  |
| 1 | 1 |  |

* + User must fill in the *prop* column with numeric values summing to 1.
* Check for the following formatting:
  + Variable named *prop*
    - Exists
    - Sums to 1 across all rows
  + Variables other than *prop* that are present in this table must also be present in the input data

## Covariate proportions

{? When individual-level data are not available, you may instead construct a simulated study population according to marginal distributions of covariates. Continuous variables are simulated assuming a truncated normal distribution specified below in **Continuous variable characteristics.** Categorical variables are simulated according to the proportional distributions specified below in **Categorical variable characteristics.**

The disadvantage of using this approach rather than individual-level data is that it assumes the covariates are independent. However, joint distributions may be specified for categorical covariates when covariance is known.}

#### ***Continuous variable characteristics***

{? Specify the distribution of values at the population level. CANTRANce will assign values to individuals by sampling from a truncated normal distribution with mean, standard deviation, minimum, and maximum values as specified in the table.

If no continuous variables are to be used, leave this section blank.}

**[Table]**

{Entry instructions: Specify a table of characteristics of distribution(s) for continuous variable(s) that describe the study population. Necessary columns are:

* *varname* – name of variable of interest. This name will be used to refer to the variable throughout the modeling process.
* *mean* – mean value in the population
* *sd* – standard deviation
* *min* – minimum allowable value
* *max* – maximum allowable value

For example, the default values specify the distribution of FIXME among individuals in the study population. Mean FIXME is FIXME with a standard deviation of FIXME. FIXME are restricted to a minimum of FIXME and a maximum of FIXME. The default values correspond to the following table: <<Show HTML example of the default>>

Note that a measure of age at study start must be included either as a continuous variable *age* in this section or as a categorical variable (either *age* or *agegroup*) in the next section.}

**${continuous\_vars}**

* For web interface:
  + Freeform string
  + Default=FIXME
* For downloadable program:
  + Populate a table with columns named *varname*, *mean*, *sd*, *min*, and *max*
  + Allow user to fill in as many rows as necessary (theoretically boundless, but we reasonably expect fewer than 10)
* Check for the following formatting:
  + Variables named *varname*, *mean*, *sd*, *min*, *max* must be present
  + No other variables allowed

#### ***Categorical variable characteristics table #XX***

{? For categorical variables, specify the proportion of individuals who fall into each covariate category in the tables below. Covariates may be specified individually, in which case independence is assumed, or as a joint distribution if covariance is known. Covariate distributions may or may not differ by trial arm.

You may specify up to five marginal or joint distributions for covariates that influence time from clinical incidence to cause-specific death.}

**[Table]**

{Entry instructions: Specify up to 5 tables that describes the proportion of individuals in the study population that fall into each covariate stratum. Column names should be the names of categorical covariates, with rows designating the covariate strata. An additional column named *prop* should provide the proportion of individuals to be assigned to each stratum.

The following categorical covariates are **required**:

* *male* – acovariate named *male* must be included in one of the five tables in order to give a distribution of the sex of individuals in the study population, with entries as 0 for female and 1 for male.
* *age* OR *agegroup* – age of individual at study start. One of these metrics must be included either as a continuous covariate (*age*) in the section above or as a categorical covariate in one of the five tables in this section.
  + If *age* is specified:
    - Use single integers, e.g. 64
  + If *agegroup* is specified:
    - Use two integers separated by a dash, e.g. 60-64
    - The model will assign a single-year age to each individual according to a uniform distribution within the age range
* *stage* – a covariate named stage must be included in one of the five tables in order to give a distribution of disease stage at clinical incidence.

Additional columns or tables may be included for covariates that have an effect on time from clinical incidence to disease-specific death.}

**${categorical\_chars1} , ${categorical\_chars2}, …, ${categorical\_chars5}**

* For web interface:
  + Freeform string
  + Default categorical\_chars1 = FIXME
  + Default categorical\_chars2 = FIXME
  + Default categorical\_chars3 = ””
  + Default categorical\_chars4 = ””
  + Default categorical\_chars5 = ””
* For downloadable program:
  + This section will have to be largely user-driven, as it’s analogous to the part where they choose the filepath for input data if they had chosen to use individual-level data rather than covariate proportions. I’d envision this section with five empty tables, or perhaps with five tables populated with a single *prop* column but not much else. The remainder of formatting depends heavily on how users choose to input their data. See below for formatting requirements.
* Check for the following formatting FOR EACH TABLE:
  + Variable named *prop*
    - Exists
    - Sums to 1
* Check for the following formatting IN ALL TABLES COMBINED (i.e. these variables need to exist in only one of the 5 input tables OR in the continuous covariates table)
  + Variable named *male*
    - Exists
    - Has values 0 AND/OR 1
  + Variable for age
    - One or more of variables named *age* and *agegroup* exists
  + Variable named *stage*
    - Exists

## Stage shift

#### ***Population stage distribution in presence of screening***

{? Specify the distribution of stages in the population when screening is present. CANTRANce will assign a stage at screen-detection based on this distribution and clinical stage for each individual. Some individuals will undergo a “stage shift” to a less advanced stage, while others will have the same stage at screen-detection and clinical incidence.

Note that these values should reflect the stage distribution *among cases that would have been clinically incident*. Thus, if overdiagnosis is a concern, the observed stage distribution in the presence of screening should be adjusted to account for overdiagnosis.}

**[Table]**

{Entry instructions: Specify the proportion of individuals in each stage in the presence of screening. Necessary columns are:

* *order* – order of stages during natural progression of the disease, with entries as 1 for the earliest, or least invasive stage of the disease, 2 for the next progression, etc.
* *stage* – stage at screen-detection. These values must match the values for clinical stage (*stage* variable) in the input data.
* *prop* – proportion of individuals in each stage

For example, the default values indicate that in a population with screening, 50% of individuals are detected in the least invasive stage, L, while 30% of individuals are in the second least invasive stage, R, and 20% of individuals are detected in the most invasive stage, D. The default values correspond to the following table: <<Show HTML example of the default>>}

**${scr\_stg\_dist}**

* For web interface:
  + Freeform string
  + Default=FIXME
* For downloadable program:
  + Populate a table with columns named *order*, *stage*, and *prop*
  + Allow user to fill in as many rows as necessary (theoretically boundless, but we reasonably expect fewer than 10)
* Check for the following formatting:
  + Variables named *order*, *stage*, and *prop* must be present
  + No other variables allowed
  + Values for stage must be the same as those in the input data *stage* variable

# CAUSE-SPECIFIC MORTALITY ESTIMATION

## Time from clinical incidence to cause-specific death

### Baseline cause-specific survival

{? CANTRANce will model time from clinical incidence to cause-specific death *in the absence of other-cause death* assuming an exponential process for the baseline cause-specific survival curve, with the option of applying modifications (e.g. relative risks or hazard ratios) for specified covariates. If you have survival data by covariate groups but not for the population as a whole, arbitrarily choose a group to serve as the baseline; statistics by covariate group will be entered in the next section.}

#### Parameter for baseline cause-specific survival curve

{? Select one survival parameter to describe cause-specific survival: k-year survival, mean survival, median survival, or a mortality rate. k-year survival refers to the proportion of people surviving the disease of interest k years following time of clinical incidence. If you choose k-year survival, you must specify the “k” below.}

**[String]**

**${mort\_param}**

* Choice string (drop-down menu)
* Allowable values:
  + mean (label: Mean survival)
  + median (label: Median survival)
  + ksurv (label: k-year survival)
  + rate (label: Mortality rate)
* Default value:
  + ksurv

IF KSURV: fill out **k (if k-year survival is chosen)**.

IF NOT KSURV: skip to **Parameter value.**

#### k (if k-year survival is chosen)

**[Numeric]**

**${mort\_k}**

* Numeric
* Default=13
* Range=[1,365] or NA

#### ***Parameter value***

{? Provide the value for the parameter selected above.

For example, the default choices set FIXME to FIXME. In other words, FIXME of individuals are expected to be alive (and FIXME of individuals are expected to have died from the disease) FIXME years after the time of incidence.

Note that this parameter should reflect cause-specific survival in the *absence* of other-cause death, generally derived from a Kaplan-Meier distribution censoring deaths from other causes. However, no such estimate was available for the example study, so cause-specific survival in the *presence* of other-cause death is used as an approximation, with the knowledge that the resultant estimate is likely inflated.}

**[Numeric]**

**${mort\_value}**

* Numeric
* Default=FIXME
* Range=TBD

### Effect of covariates on cause-specific survival

{? If covariates modify the risk of cause-specific death from the baseline cause-specific survival curve specified above, you may specify the effect of the covariates here as either: 1) survival statistics for the same parameter (ksurv, mean, median, rate) chosen for the baseline curve, or 2) hazard ratios for direct comparison to the baseline curve.

Any covariate in the data may modify incidence-free survival; they may be treated as categorical variables or continuous variables. If cause-specific survival varies by age of individual, age may be used either as a continuous variable *age* (e.g. 64) or as a categorical variable *agegroup* (e.g. 60-64), regardless of the way in which age was specified in the input data.

In the first table, specify how stage at incidence modifies the risk of cause-specific death. You may specify up to two additional covariates. For each, specify the survival statistics or hazard ratios in the corresponding table below.}

#### Covariate #XX: Effect on cause-specific survival

**[Table]**

**${mort\_covar1}, ${mort\_covar2}, ${mort\_covar3}**

{Entry instructions: Specify a 2- or 3-column table that describes the effect of the chosen covariate on cause-specific survival. For each table, 1 covariate and 1 parameter variable must be specified.

Covariate variable:

* covariate – name this column with the name of the covariate of interest. For a categorical covariate, the entries should be the category values. For a continuous covariate, the entries should be NA.

Parameter variable (PICK ONE):

* *stat* (applicable for categorical covariates only) – value for survival statistic chosen for the baseline curve. For example, if mortality rate was chosen as the parameter for the baseline curve, entries in this column should correspond to the mortality rate for each covariate group.
* *HR* – hazard ratio for cause-specific survival **rate** for each covariate stratum compared to the baseline **rate** (note that the ratio should be a ratio of rates). If one of the groups corresponds to the baseline group for a categorical covariate, enter a hazard ratio of 1. For a continuous covariate, provide the hazard ratio associated with each unit increase in the value of the covariate.

For example, the default values correspond to the following table, which indicates FIXME ranging from FIXME to FIXME depending on FIXME: <<Show HTML example of the default>>}

* For web interface:
  + Freeform string
  + Default mort\_covar1 = FIXME
  + Default mort\_covar2 = “”
  + Default mort\_covar3 = “”
* For downloadable program:
  + Force selection of *stage*, and prompt user to select up to 2 additional covariates from among the variables in the input data
  + For covariates that have been chosen (what’s the best way to do this in the interface?):
    - If the covariate is categorical (as chosen by the drop-down menu currently), they may choose whether they will input a *stat* value or a *HR* value in the table, and a corresponding (empty) column should be added to the table. If it’s continuous, an empty *HR* value should be added automatically (no choice between *stat* and *HR*).
* Check for the following formatting:
  + Metric variable
    - Either *stat* or *HR* exists, but not both
  + Variables in these tables (other than *stat* and *HR*) must be present in the input data

# OTHER-CAUSE MORTALITY ESTIMATION

## Time from study start to other-cause death

#### Hazard ratio for other-cause death compared to general population

{? For a typical population, time to other-cause death is approximated using US cohort life tables matched to each individual by sex and year of birth. If the study population is thought to be healthier or less healthy than the US population, the user can specify a hazard ratio to adjust the life table-based survival estimates. This input may be left blank.}

**[Numeric]**

{Entry instructions: Enter a hazard ratio for other-cause death from time of clinical incidence.

For example, a value of 0.7 indicates that the study population is expected to have a risk of other-cause death 70% that of the US population as a whole.}

**${ocd\_HR}**

* Numeric
* Default=1
* Range=[0,20]