README

This folder contains the necessary code to compute the various validation results for the updated PANGEA trajectory models (BM and no BM) and alternatives (original PANGEA and 20/2/20).

The folder includes:

- validation_code.R: main script to run to perform the analysis
- data_preparation.R: script computing the trajectories, risk scores, and some other variables needed for the analysis
- table_descriptives.R: script to compute the basic descriptive statistics of the cohort (sourced in validation_code.R)
- risk_reclassification.R: script to compute risk strata from the PANGEA risk scores (sourced in validation_code.R)
- Functions.R: script defining the functions needed to run the various analyses
- Packages.R: script loading the needed packages
- cutoffs_pangea_training.rda: file storing cutoffs used to determine risk strata
- models: folder storing the PANGEA models and necessary information to compute the risk predictions
- results: folder to save the results at the end of the analysis (currently empty)

How to perform the analysis

- Set the "Validation - Code for External Collaborators" folder as working directory
- Open the R script "validation_code.R"
- Modify the path to load your data (at the beginning of the script)
- If applicable, on line 7 set save_deidentified to TRUE (instead of FALSE, the default value)
- Run the "validation_code.R" script

The results will be saved in the results folder.

The validation_code.R executes the following steps:

1. Source the Functions.R and Packages.R scripts
2. Load the data
3. Source the data_preparation.R script to compute some variables needed for the analysis, including the trajectories, the rolling 20/2/20 classification, and PANGEA risk scores (including for some previous developmental versions of the model)
    (a) **Note**: To just compute the PANGEA 2.0 risk scores, simply run validation_code.R up to line 32 where data_preparation.R is sourced.
    (b) The PANGEA 2.0 models are saved in the pangea_bm_traj_alt.RData and pangea_no_bm_traj_alt.RData files in the models folder, and their risk score variables created by data_preparation.R have similar names.

4. Compute descriptive statistics for the cohort
5. Compute basic statistics on progression rates for the cohort
6. Compute descriptive statistics on the risk scores for the cohort
7. Compute concordance statistics (ability of the risk scores to appropriately rank patients)
8. Compute classification statistics (based on risk strata)
9. Compute calibration statistics (ability of risk scores to match the actual progression risk for various subgroups)
10. Store the results in the results folder

Format of the data

The dataset has to include the following variables:

tstart: Time at which the current labs are measured (it should be 0 for the first observation) in years; used for validation.

tstop: Time of the next lab or the end of the follow up (final observation) in years; used for validation.

prog_mm: Binary progression status at the time tstop (1 = slim-crab myeloma, 0 otherwise); used for validation.

mspike: M-spike in g/dL; used for the PANGEA risk prediction.

iuratio: Involved/Uninvolved free light chain ratio; used for the PANGEA risk prediction.

plasmacells: Numeric percentage of plasma cells (0-100); used for the PANGEA risk prediction.

creatinine: Creatinine in mg/dL; used for the PANGEA risk prediction.

age: Age of the patient at each visit in years; used for the PANGEA risk prediction.

hgb: Hemoglobin in g/dL (needed to compute the trajectory); used for the PANGEA risk prediction.

sex: Sex of the patient; used only for descriptive statistics.

current_diagnosis: Categorical variable, is the patient smoldering, i.e. "SMM" (for this analysis we want only smoldering patients); used only for descriptive statistics.

end_date_type: Description of the event that caused censoring or progression. possible values for censored patients are "Death", "Last appointment", and "Other treatment". For patients who progress the value should just be "Diagnosis". Used only for descriptive statistics.

race: Race of patient. Possible values are "Asian", "Black or African American", "Declined", "Multiple", "Other", and "White". For patients where race is missing (it's ok if that's everyone in your cohort), the value should be NA. Used only for descriptive statistics.

ethnicity: Ethnicity of the patient. Possible values are "Declined", "Hispanic or Latino", and "Not Hispanic or Latino". For patients where ethnicity is missing (it's ok if that's everyone in your cohort), the value should be NA. Used only for descriptive statistics.

immunofix2: Immunofixation isotype. Possible values are "Biclonal", "IgA", "IgG", and "Light Chain Only". Missing values should be indicated as NA. Used only for descriptive statistics.

The data has to be in a long format, which is, each patient visit corresponds to one row in the dataset.

Note: PANGEA risk predictions can only be computed for the rows in which all predictor variables (mspike, iuratio, creatinine, hgb, and age) are available (not NA). Because of this, labs associated with the same patient visit but reported on slightly different dates (e.g. mspike on 5/9/24 and iuratio on 5/4/24) should be combined into a single row with a single date in order to get a risk prediction for that visit.

For a visual aide, please see the first rows of our dataset, obtained using the R function `tmerge` from the package `survival`. The data are loaded in data.table format.

```
## Dataset
library(data.table)
sample_data <- fread("fake_example_data.csv")
sample_data[participant_id == 1,]
```

```
##       V1 participant_id    tstart     tstop prog_mm mspike   iuratio
##    <int>          <int>     <num>     <num>   <int>  <num>     <num>
## 1:     1              1 0.0000000 0.9117243       0   0.37  7.903827
## 2:     2              1 0.9117243 1.4100240       0   1.93 17.548807
## 3:     3              1 1.4100240 2.7132696       0   1.58 19.952914
## 4:     4              1 2.7132696 3.2745413       0   1.39 67.277014
##    plasmacells      age creatinine   hgb current_diagnosis    end_date_type
##          <int>    <num>      <num> <num>            <char>           <char>
## 1:          NA 42.96655       0.44  12.0               SMM Last appointment
## 2:          NA 43.87828       1.15  10.2               SMM Last appointment
## 3:          NA 44.37658       0.74  10.2               SMM Last appointment
## 4:          NA 45.67982       0.34  10.4               SMM Last appointment
##                         race          ethnicity immunofix2    sex
##                       <char>             <char>     <char> <char>
## 1: Black or African American Hispanic or Latino        IgA   Male
## 2: Black or African American Hispanic or Latino        IgA   Male
## 3: Black or African American Hispanic or Latino        IgA   Male
## 4: Black or African American Hispanic or Latino        IgA   Male
```

```
sample_data[participant_id == 2,]
```

```
##       V1 participant_id    tstart     tstop prog_mm mspike  iuratio plasmacells
##    <int>          <int>     <num>     <num>   <int>  <num>    <num>       <int>
## 1:     5              2 0.0000000 0.2956944       0   2.00       NA          NA
## 2:     6              2 0.2956944 0.5256789       0   2.50 59.73815          45
## 3:     7              2 0.5256789 0.8076837       0   4.11 78.20539          45
## 4:     8              2 0.8076837 1.3032455       0   3.69 95.81232          45
## 5:     9              2 1.3032455 1.4948993       1   3.85 91.67899          45
##         age creatinine   hgb current_diagnosis end_date_type   race ethnicity
##       <num>      <num> <num>            <char>        <char> <char>    <char>
## 1: 77.00214       0.56   9.8               SMM     Diagnosis  White      <NA>
## 2: 77.29784       0.17   8.4               SMM     Diagnosis  White      <NA>
## 3: 77.52782       0.27  10.1               SMM     Diagnosis  White      <NA>
## 4: 77.80983       0.72  10.0               SMM     Diagnosis  White      <NA>
## 5: 78.30539       1.19  10.9               SMM     Diagnosis  White      <NA>
##    immunofix2    sex
##        <char> <char>
## 1:        IgG Female
## 2:        IgG Female
## 3:        IgG Female
## 4:        IgG Female
## 5:        IgG Female
```

R session

As a reference, we report the R version we used

```
R.Version()$version.string
```

```
## [1] "R version 4.4.2 (2024-10-31)"
```