

expectedCases: Projecting Cancer Incidence

Overview

The **expectedCases** package estimates projected cancer incidence and mortality for a study cohort, given age-specific incidence and mortality rates. This vignette walks through a reproducible workflow with example data, explanations, and simple visualization.

What this vignette shows.

How to: (1) prepare a cohort in 5-year age bands, (2) supply sex-specific incidence and mortality rates, (3) run the projection, and (4) interpret/visualize results.

Cohort requirement

Step 1. Create the baseline cohort

Each age band must represent a **5-year range** (e.g., “40–44”, “45–49”, etc.). The function assumes equally spaced 5-year intervals for age expansion.

```
counts_5y <- tibble::tibble(
  age_band = c("40-44", "45-49", "50-54", "55-59"),
  N         = c(2694, 3480, 4166, 4964)
)
counts_5y
#> # A tibble: 4 × 2
#>   age_band      N
#>   <chr>      <dbl>
#> 1 40-44    2694
#> 2 45-49    3480
#> 3 50-54    4166
#> 4 55-59    4964
```

Step 2. Define incidence and mortality rates

Rates are specified separately for females and males.

- **Incidence rates** are typically provided in *broader grouped age bands* (e.g., “0–24”, “25–39”, “40–54”, “55–69”, “70–85+”).

- **Mortality rates**, by contrast, are usually specified in *detailed 5-year bands* (e.g., “40–44”, “45–49”, “50–54”, etc.).

Note that while the cohort (counts_5y) must always use 5-year age intervals (e.g., “40–44”, “45–49”), the incidence tables may use broader age groups (e.g., “40–54”). The function automatically splits these wider bands into matching 5-year intervals internally for calculation.

```
library(expectedCases)
library(tibble)

# Sex proportions
female_share <- 0.6
male_share   <- 0.4

# Incidence and mortality rates per 100k (simplified integers)
female_inc_wide <- tibble(
  band = c("40-54", "55-69", "70-85"),
  rate_per100k = c(39, 102, 278)
)
female_inc_wide
#> # A tibble: 3 × 2
```

```

#>   band   rate_per100k
#>   <chr>         <dbl>
#> 1 40-54           39
#> 2 55-69          102
#> 3 70-85          278

female_mort_5y <- tibble(
  band = c("40-44", "45-49", "50-54", "55-59", "60-64", "65-69", "70-74", "75-79", "80-84", "85+" ),
  rate_per100k = c(87, 138, 214, 309, 455, 707, 1164, 2104, 4133, 13700)
)
female_mort_5y
#> # A tibble: 10 × 2
#>   band   rate_per100k
#>   <chr>         <dbl>
#> 1 40-44           87
#> 2 45-49          138
#> 3 50-54          214
#> 4 55-59          309
#> 5 60-64          455
#> 6 65-69          707
#> 7 70-74          1164
#> 8 75-79          2104
#> 9 80-84          4133
#> 10 85+          13700

male_inc_wide <- tibble(
  band = c("40-54", "55-69", "70-85"),
  rate_per100k = c(45, 148, 338)
)
male_inc_wide
#> # A tibble: 3 × 2
#>   band   rate_per100k
#>   <chr>         <dbl>
#> 1 40-54           45
#> 2 55-69          148
#> 3 70-85          338

male_mort_5y <- tibble(
  band = c("40-44", "45-49", "50-54", "55-59", "60-64", "65-69", "70-74", "75-79", "80-84", "85+" ),
  rate_per100k = c(155, 225, 337, 506, 789, 1190, 1888, 3239, 5974, 15415)
)
male_mort_5y
#> # A tibble: 10 × 2
#>   band   rate_per100k
#>   <chr>         <dbl>
#> 1 40-44          155
#> 2 45-49          225
#> 3 50-54          337
#> 4 55-59          506
#> 5 60-64          789
#> 6 65-69          1190
#> 7 70-74          1888
#> 8 75-79          3239
#> 9 80-84          5974
#> 10 85+          15415

```

Step 3. Run the projection

Function Arguments

| Argument | Type | Description |
|--------------|---------|--|
| counts_5y | tibble | Baseline cohort counts by 5-year age bands (e.g., “40-44”, “45-49”), with one column <code>age_band</code> and one column <code>N</code> (population size). |
| female_share | numeric | Proportion of the total cohort that is female (between 0 and 1). |

| Argument | Type | Description |
|-----------------|----------------|---|
| male_share | numeric | Proportion of the total cohort that is male (between 0 and 1). Must satisfy $\text{female_share} + \text{male_share} = 1$. |
| female_inc_wide | tibble | Female incidence rates per 100,000 population, specified in broader age bands (e.g., “40-54”, “55-69”, “70-85”). The function interpolates these to 5-year intervals internally. |
| female_mort_5y | tibble | Female all-cause mortality rates per 100,000 population, given in 5-year bands (e.g., “40-44”, “45-49”, ..., “85+”). |
| male_inc_wide | tibble | Male incidence rates per 100,000 population, formatted like <code>female_inc_wide</code> . |
| male_mort_5y | tibble | Male all-cause mortality rates per 100,000 population, formatted like <code>female_mort_5y</code> . |
| study_start | numeric | Starting calendar year of the cohort (e.g., 2020). |
| study_end | numeric | Ending year of data collection or baseline projection period (e.g., 2025). Used to determine midpoint for internal timing. |
| age_min | numeric | Minimum age of the modeled cohort (default: 40). |
| age_max | numeric | Maximum modeled age (default: 100). Individuals exceeding this are dropped from the cohort. |
| end_year | numeric | Final projection year (e.g., 2040). Determines length of simulation. |
| begin_year | numeric | Optional. Custom first projection year (default = midpoint of <code>[study_start, study_end]</code>). |
| diag_years | numeric vector | Optional. Specific years at which projections (e.g., 2022, 2030, 2040) are returned in the output. |
| print_markdown | logical | If TRUE, prints a formatted markdown summary table of expected cases in the console. Set FALSE for silent operation (recommended for vignettes). |

Return Value

The function returns a **list** with two tibbles:

| Output | Description |
|----------------|---|
| summary_tbl | Aggregated expected cancer cases by sex and projection period. |
| projection_tbl | Detailed year-by-year projection showing <code>alive_start</code> , <code>new_cases_year</code> , <code>deaths_year</code> , <code>aged_out_year</code> , <code>alive_end</code> , and <code>cum_cases</code> . |

```
result <- run_cancer_projection(  
  counts_5y,  
  female_share, male_share,  
  female_inc_wide, female_mort_5y,  
  male_inc_wide, male_mort_5y,  
  study_start = 2020, study_end = 2025,
```

```

age_min = 40, age_max = 100,
end_year = 2040,
print_markdown = FALSE
)

```

Statistical Framework

The population projection in `expectedCases` follows a deterministic cohort-based model, where each age–sex group evolves over discrete yearly time steps.

For each age group a , sex s , and year t :

$$\begin{aligned}
 \text{NewCases}_{a,s,t} &= N_{a,s,t} \times \lambda_{a,s,t}, \\
 \text{Deaths}_{a,s,t} &= N_{a,s,t} \times \mu_{a,s,t}, \\
 \text{AliveEnd}_{a,s,t} &= N_{a,s,t} - \text{NewCases}_{a,s,t} - \text{Deaths}_{a,s,t}, \\
 N_{a+1,s,t+1} &= \text{AliveEnd}_{a,s,t}.
 \end{aligned}$$

where:

| Symbol | Meaning |
|---------------------------|--|
| $N_{a,s,t}$ | Individuals alive and cancer-free at the start of year t . |
| $\lambda_{a,s,t}$ | Age–sex–specific incidence rate (probability of developing cancer). |
| $\mu_{a,s,t}$ | Age–sex–specific mortality rate (probability of death from any cause). |
| $\text{NewCases}_{a,s,t}$ | Expected new cancer cases during year t . |
| $\text{Deaths}_{a,s,t}$ | Expected all-cause deaths during year t . |
| $\text{AliveEnd}_{a,s,t}$ | Number of individuals surviving cancer-free to the end of year t . |

Continuous-Time Analogy

In continuous time, the discrete recursion approximates the differential equation:

$$\frac{dN_{a,s}(t)}{dt} = -N_{a,s}(t) \times (\lambda_{a,s}(t) + \mu_{a,s}(t)).$$

which represents competing risks between cancer incidence and mortality.

In the discrete (annual) version used here, the yearly transition is approximated as:

$$N_{a+1,s,t+1} \approx N_{a,s,t} \times [1 - (\lambda_{a,s,t} + \mu_{a,s,t})].$$

Step 4. Interpret the results

The output includes:

- `summary_tbl`: summarized expected cases by sex and projection period
- `yearly_tbl`: detailed year-by-year estimates (alive, new cases, deaths, cumulative cases)

```

result$summary_tbl
#> # A tibble: 3 × 3
#>   Sex      `Time Period` `Expected Cases`
#>   <chr>    <chr>          <dbl>
#> 1 Female 2022–2040      178.
#> 2 Male   2022–2040      155.
#> 3 Total  2022–2040      333.

result$projection_tbl
#> # A tibble: 6 × 8
#>   Sex      year alive_start new_cases_year deaths_year aged_out_year alive_end
#>   <chr>   <int>      <dbl>         <dbl>      <dbl>         <dbl>    <dbl>
#> 1 Female 2022      9182.         5.46       18.8          0      9164.
#> 2 Female 2030      8980.         7.63       35.6          0      8944.
#> 3 Female 2040      8447.        15.6       85.6          0      8361.
#> 4 Male   2022      6122.         4.80       20.5          0      6101.
#> 5 Male   2030      5899.         7.08       39.2          0      5860.

```

```
#> 6 Male 2040 5333. 12.6 86.0 0 5247.
#> # 1 more variable: cum_cases <dbl>
```

| Projection | | | | | | | |
|------------|------|-------------|----------------|-------------|---------------|-----------|-----------|
| Sex | year | alive_start | new_cases_year | deaths_year | aged_out_year | alive_end | cum_cases |
| Female | 2022 | 9182.4 | 5.5 | 18.8 | 0 | 9163.6 | 5.5 |
| Female | 2030 | 8979.7 | 7.6 | 35.6 | 0 | 8944.1 | 59.4 |
| Female | 2040 | 8447.1 | 15.6 | 85.6 | 0 | 8361.5 | 177.5 |
| Male | 2022 | 6121.6 | 4.8 | 20.5 | 0 | 6101.1 | 4.8 |
| Male | 2030 | 5898.7 | 7.1 | 39.2 | 0 | 5859.5 | 54.1 |
| Male | 2040 | 5332.9 | 12.6 | 86.0 | 0 | 5246.9 | 155.2 |

Description of projection table columns in `result$projection_tbl`.

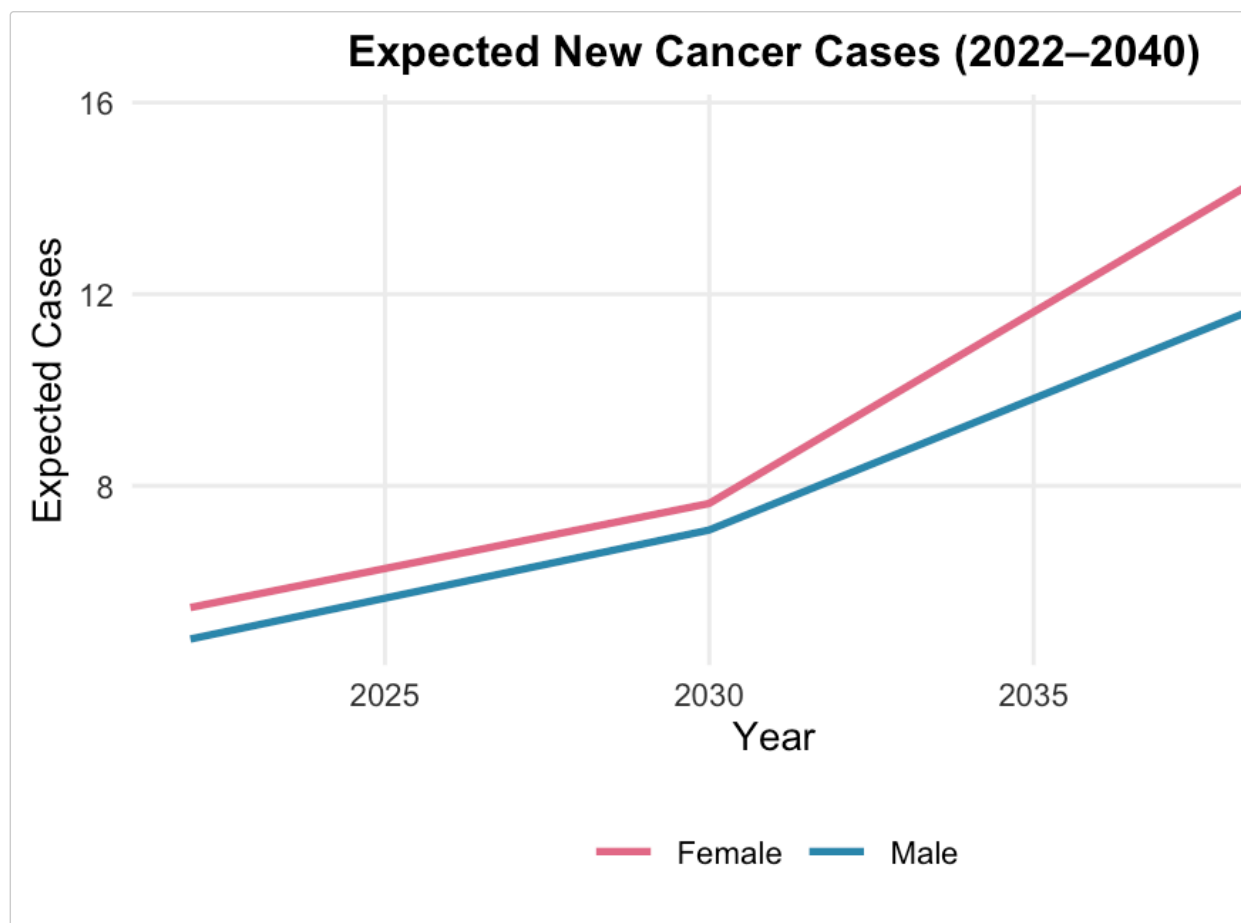
| Column | Description |
|----------------|---|
| Sex | Sex of the subgroup (Female or Male). |
| year | Calendar year of projection. |
| alive_start | Number alive at the start of the year within modeled ages. |
| new_cases_year | Expected new cancer cases during that year. |
| deaths_year | Expected all-cause deaths during that year. |
| aged_out_year | Individuals reaching <code>age_max</code> who exit the modeled range next year. |
| alive_end | Number of survivors remaining within modeled ages at year-end. |
| cum_cases | Cumulative expected cancer cases up to and including that year. |

Step 5. Visualize trends

We can plot the expected cumulative cases over time by sex.

```
library(ggplot2)

ggplot(result$projection_tbl, aes(x = year, y = new_cases_year, color = Sex)) +
  geom_line(linewidth = 0.9) +
  scale_color_manual(values = c("Female" = "#E16A86", "Male" = "#2E86AB")) + # soft pink &
    teal-blue
  labs(
    title = "Expected New Cancer Cases (2022–2040)",
    x = "Year",
    y = "Expected Cases"
  ) +
  theme_minimal(base_size = 10) + # smaller text
  theme(
    legend.position = "bottom", # legend below for balance
    legend.title = element_blank(),
    axis.text = element_text(color = "gray20"),
    plot.title = element_text(face = "bold", size = 11, hjust = 0.5),
    panel.grid.minor = element_blank()
  )
```



Help

- For help and documentation:

`?run_cancer_projection`

Session information for reproducibility

`sessionInfo()`