

# expectedCases: Projecting Cancer Incidence

## Contents

<b>1</b>	<b>expectedCases Tutorial</b>	<b>1</b>
1.1	Overview . . . . .	1
1.2	Cohort requirements . . . . .	1
1.3	Step 1. Create the baseline cohort . . . . .	1
1.4	Step 3. Run the projection . . . . .	3
1.5	Function Arguments . . . . .	3
1.6	Statistical Framework . . . . .	4
1.7	Step 4. Interpret the results . . . . .	5
1.8	Step 5. Visualize trends . . . . .	7
1.9	Help . . . . .	7
1.10	Session information for reproducibility . . . . .	7

## 1 expectedCases Tutorial

### 1.1 Overview

The **expectedCases** package projects expected cancer incidence and mortality for a cohort using age-specific rates. This vignette shows how to (1) prepare a 5-year age-band cohort, (2) supply sex-specific incidence and mortality, (3) run projections, and (4) interpret/visualize results.

### 1.2 Cohort requirements

Cohort counts must be in **5-year age bands** (e.g., 40-44, 45-49, ...). Incidence may be in broader bands (e.g., 40-54, 55-69); the function maps those to single years internally. Open-ended bands (e.g., 85+) are supported and extended to `age_max`.

### 1.3 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 x 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.

```
# Sex proportions
female_share <- 0.6
male_share   <- 0.4

# Incidence and mortality rates per 100k (simplified integers)
# IMPORTANT: Include an open-ended "85+" band so coverage reaches age_max (e.g., 100)

female_inc_wide <- tibble(
  band = c("40-54", "55-69", "70-84", "85+"),
  rate_per100k = c(39, 102, 278, 400) # last value is illustrative; use your preferred rate
)

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)
)

male_inc_wide <- tibble(
  band = c("40-54", "55-69", "70-84", "85+"),
  rate_per100k = c(45, 148, 338, 500) # last value is illustrative; use your preferred rate
)

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)
)
```

**Validation:** Age bands must be digits only (NN-NN or NN+). Typos like 4o-44 (letter o) are rejected with a clear error. Rates must cover **every** age in age\_min:age\_max exactly once (no gaps/overlaps).

## 1.4 Step 3. Run the projection

### 1.5 Function Arguments

Argument	Type	Description
counts_5y	tibble	Baseline cohort counts by <b>5-year age bands</b> (e.g., “40-44”, “45-49”), with one column age_band and one column N (population size).
female_share	numeric	Proportion of the total cohort that is female (between 0 and 1).
male_share	numeric	Proportion of the total cohort that is male (between 0 and 1). Must satisfy female_share + male_share = 1.
female_inc_wide	tibble	Female <b>incidence rates</b> per 100,000 population, specified in broader age bands (e.g., “40-54”, “55-69”, “70-85”). The function maps these broader age bands to single-year values internally, assuming flat rates within each band.
female_mort_5y	tibble	Female <b>all-cause mortality rates</b> 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 female_inc_wide.
male_mort_5y	tibble	Male all-cause mortality rates per 100,000 population, formatted like female_mort_5y.
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 [study_start, study_end]).
diag_years	numeric vector	Optional. Specific years at which projections (e.g., 2022, 2030, 2040) are returned in the output.

Argument	Type	Description
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 alive_start, new_cases_year, deaths_year, aged_out_year, alive_end, and cum_cases.

```
res <- run_cancer_projection(
  counts_5y,
  female_share = female_share, male_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,
  diag_years = 2022:2040,
  print_markdown = FALSE
)

projection_tbl <- res$projection_tbl
summary_tbl <- res$summary_tbl
```

Before interpreting results, it may help to understand the model structure used internally by `expectedCases`.

## 1.6 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).
NewCases $_{a,s,t}$	Expected new cancer cases during year $t$ .
Deaths $_{a,s,t}$	Expected all-cause deaths during year $t$ .
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 discretizes the continuous-time hazard formulation, which represents competing risks between cancer incidence and mortality.

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

which represents competing risks between cancer incidence and mortality.

In the discrete (annual) version used here, the yearly transition follows:

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

## 1.7 Step 4. Interpret the results

The output includes:

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

```
knitr::kable(summary_tbl, caption = "Expected cases by sex and period")
```

Table 4: Expected cases by sex and period

Sex	Time Period	Expected Cases
Female	2022-2040	177.5
Male	2022-2040	155.2
Total	2022-2040	332.7

```
projection_tbl |>
  dplyr::select(Sex, year, alive_start, new_cases_year, deaths_year, aged_out_year, alive_end, cum_cases)
  dplyr::mutate(dplyr::across(where(is.double), ~ round(.x, 1))) |>
  knitr::kable(caption = "Year-by-year projection (rounded for display)")
```

Table 5: Year-by-year projection (rounded for display)

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	2023	9163.6	5.8	20.6	0	9142.9	11.2
Female	2024	9142.9	6.1	22.4	0	9120.6	17.3
Female	2025	9120.6	6.4	24.1	0	9096.4	23.6
Female	2026	9096.4	6.6	25.9	0	9070.6	30.3
Female	2027	9070.6	6.9	27.6	0	9043.0	37.2
Female	2028	9043.0	7.2	30.3	0	9012.7	44.4
Female	2029	9012.7	7.4	33.0	0	8979.7	51.8
Female	2030	8979.7	7.6	35.6	0	8944.1	59.4
Female	2031	8944.1	7.9	38.3	0	8905.8	67.3
Female	2032	8905.8	8.1	40.9	0	8864.9	75.3
Female	2033	8864.9	9.2	45.3	0	8819.6	84.6
Female	2034	8819.6	10.4	49.7	0	8770.0	94.9
Female	2035	8770.0	11.5	54.0	0	8716.0	106.4
Female	2036	8716.0	12.5	58.2	0	8657.8	118.9
Female	2037	8657.8	13.6	62.3	0	8595.5	132.6
Female	2038	8595.5	14.3	70.3	0	8525.2	146.9
Female	2039	8525.2	15.0	78.1	0	8447.1	161.8
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	2023	6101.1	5.1	22.5	0	6078.6	9.9
Male	2024	6078.6	5.4	24.6	0	6054.0	15.4
Male	2025	6054.0	5.8	26.6	0	6027.5	21.1
Male	2026	6027.5	6.1	28.5	0	5999.0	27.2
Male	2027	5999.0	6.4	30.5	0	5968.5	33.5
Male	2028	5968.5	6.6	33.4	0	5935.0	40.1
Male	2029	5935.0	6.8	36.3	0	5898.7	47.0
Male	2030	5898.7	7.1	39.2	0	5859.5	54.1
Male	2031	5859.5	7.3	42.0	0	5817.6	61.4
Male	2032	5817.6	7.5	44.7	0	5772.9	68.9
Male	2033	5772.9	8.4	49.1	0	5723.7	77.3
Male	2034	5723.7	9.2	53.5	0	5670.3	86.4
Male	2035	5670.3	10.0	57.6	0	5612.6	96.4
Male	2036	5612.6	10.7	61.7	0	5550.9	107.1
Male	2037	5550.9	11.5	65.7	0	5485.3	118.5
Male	2038	5485.3	11.9	72.8	0	5412.5	130.4
Male	2039	5412.5	12.2	79.6	0	5332.9	142.6
Male	2040	5332.9	12.6	86.0	0	5246.9	155.2

Table 6: Description of projection table columns in `projection_tbl`.

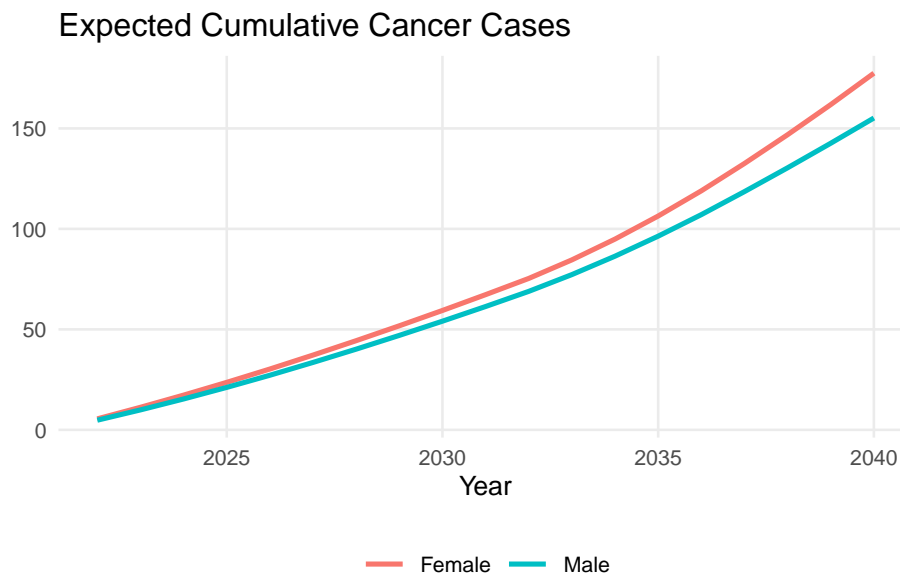
Column	Description
Sex	Sex of the subgroup (Female or Male).
year	Calendar year of projection.
alive_start	Number alive at the start of the year within modeled ages.
new_cases_year	Expected new cancer cases during that year.
deaths_year	Expected all-cause deaths during that year.
aged_out_year	Individuals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

Column	Description
cum_cases	Cumulative expected cancer cases up to and including that year.

## 1.8 Step 5. Visualize trends

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

```
ggplot(projection_tbl, aes(year, cum_cases, color = Sex)) +
  geom_line(linewidth = 0.9) +
  labs(title = "Expected Cumulative Cancer Cases", x = "Year", y = "") +
  theme_minimal(base_size = 10) +
  theme(
    legend.position = "bottom",
    legend.title = element_blank(),
    panel.grid.minor = element_blank()
  )
```



## 1.9 Help

- For help and documentation:

```
?run_cancer_projection
```

## 1.10 Session information for reproducibility

```
sessionInfo()
```