Projecting cancer prevalence through combining age-period-cohort models and survival models.

Paul C. Lambert[1,2]

[1]Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway
[2]Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

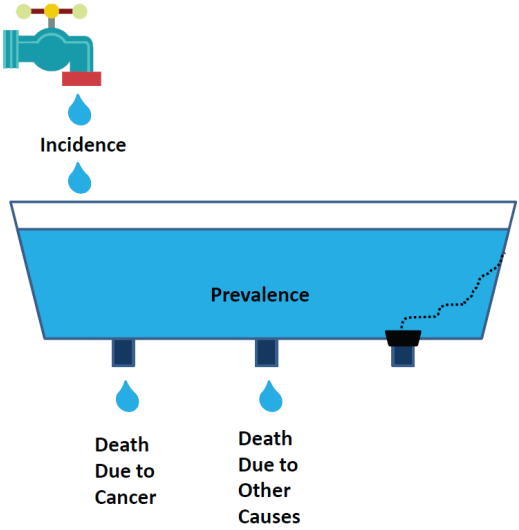
Northern European Stata Conference, Oslo
2026-09-24

Background

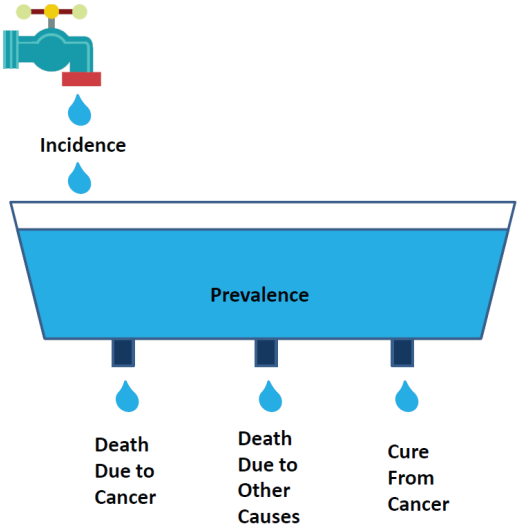
- Important to investigate and understand patterns and trends in incidence/mortality rates.
- Need to estimate the number of future cases for planning of treatment and care.
- Assuming that the current rates will remain the same is often inadequate.
- We are interested in understanding:

Incidence	The number (rate) of new cancer cases
Survival	The probability of surviving after a diagnosis of cancer
Prevalence	The number (proportion) of people living with a diagnosis of cancer.

Prevalence


Incidence
Prevalence
Death
Due to
Cancer
Death
Due to
Other
Causes

Prevalence


Incidence
Prevalence
Death Due to Cancer
Death Due to Other Causes
Cure From Cancer

Prevalence: Counting Method

- Cancer registries that have collected data for a long time, can calculate prevalence by simply counting the number of people who are alive at a specific calendar period.
- For x-year prevalence we count how many people are alive *and* have been diagnosed within the previous x years.
- For younger cancer registries prevalence needs to be estimated using modelling as there are details of individuals diagnosed prior to the start of the registry[1].
  - The Cancer Registry of Norway starting collecting data in 1952.
- To predict future prevalence we need to model what happens in the future.

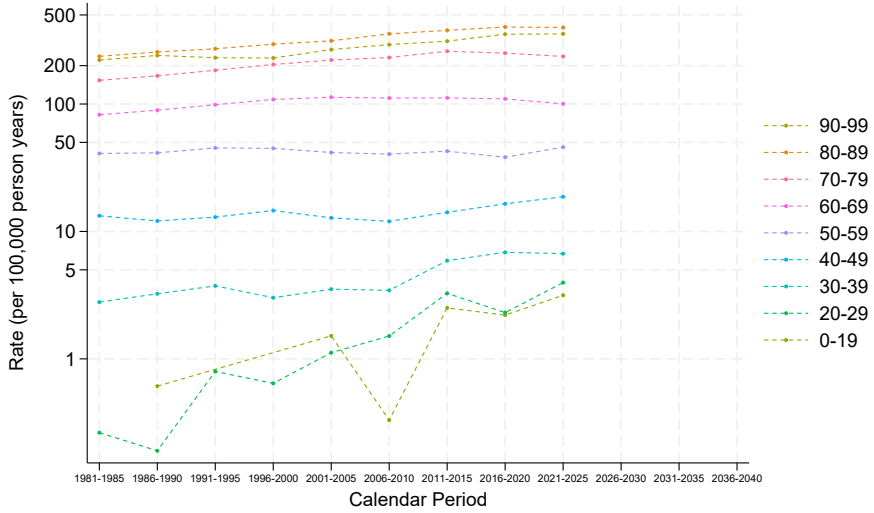
Prevalence is a function of incidence and survival

- Future cancer prevalence is dependent on
  - Current prevalence
  - Future incidence
  - Future survival
  - Future population changes

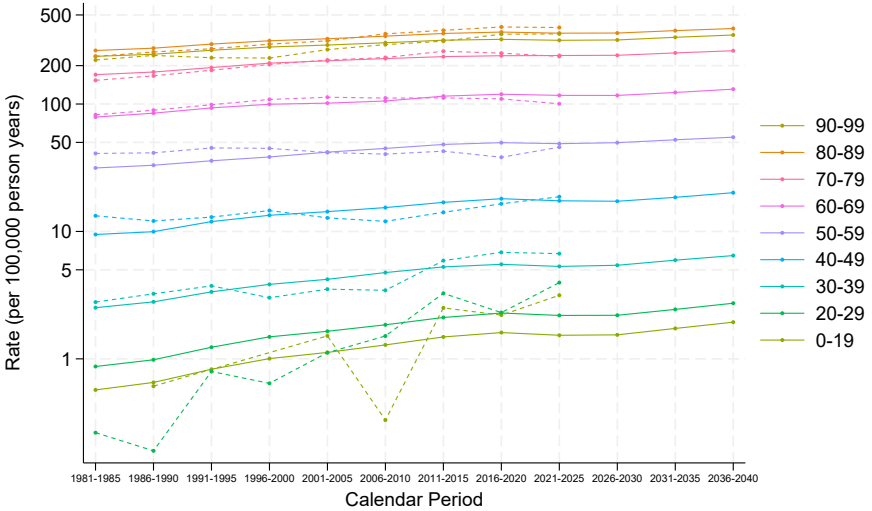
Prevalence is a function of incidence and survival

- Future cancer prevalence is dependent on
  - Current prevalence
  - Future incidence
  - Future survival
  - Future population changes
- We need to model incidence: Age Period Cohort (APC) models
- We need to model survival: Flexible Parametric Survival Model
- We need to combine the two models and incorporate future population changes.
- We want to see how future prevalence is affected by different assumptions about future incidence and survival.

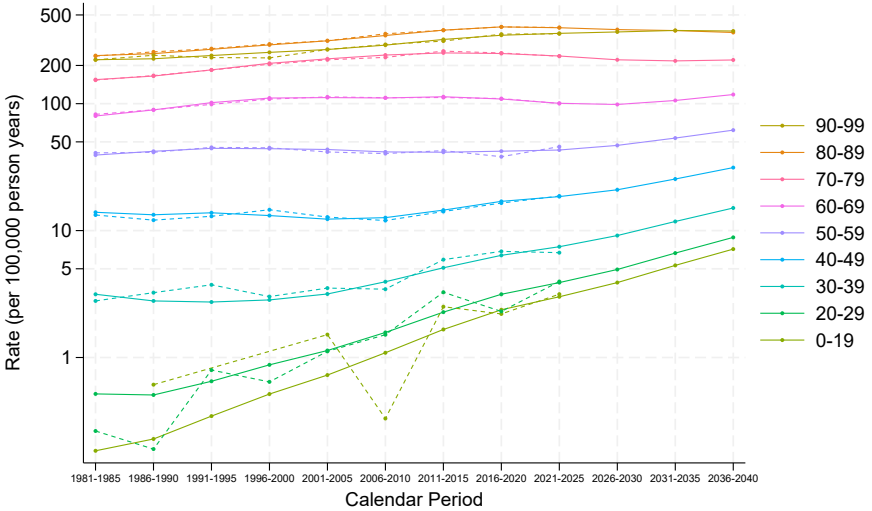
Female Colon Cancer: Modelling Incidence (Observed)



Female Colon Cancer: Modelling Incidence (Age Period Model)



Female Colon Cancer: Modelling Incidence (Age Period Cohort Model)



Age Period and Age Cohort models

- The Age-Period (AP) model is

$$\log(\text{rate}) = f(a) + g(p)$$

$f(a)$ is a function of age and $g(p)$ is a function of calendar period (of diagnosis).

- The Age-Cohort (AC) model is

$$\log(\text{rate}) = f(a) + h(c)$$

$f(a)$ is a function of age and $h(c)$ is a function of birth cohort.

- Spline functions are used to estimate $f(a)$, $g(p)$ and $h(c)$, similar to other implementations[2, 3, 4].
- In both models we can estimate the linear effect of calendar time, which is known as the drift parameter.

Age Period Cohort (APC) model

- Neither the Age-Period or Age-Cohort models may fit the data well.
- Can we fit a model with the three terms simultaneously?

Age Period Cohort (APC) model

- Neither the Age-Period or Age-Cohort models may fit the data well.
- Can we fit a model with the three terms simultaneously?
- We have an identifiability problem: age, period and cohort are linearly related.

$$\text{Age} = \text{Period} - \text{Cohort}$$

- Attempts to 'overcome' the identifiability problem, but need to apply constraints in order to estimate separate functions of age, period and cohort[2, 5].

Age period cohort models: constraints

- The APC model is

$$\log(\text{rate}) = f(a) + g(p) + h(c)$$

  $f(a)$ is a function of age, $g(p)$ is a function of calendar period, and $h(c)$ is a function of birth cohort.
- The identifiability issue leads to too many parameters and we cannot estimate them all.
- Need to apply constraints in order to plot separate functions of age, period and cohort.

Age period cohort models: constraints

- The absolute (log) rates are estimated for age at a given reference period/cohort.
- Using matrix transformations of the design matrices, it is possible to fully extract the linear trend of $g(p)$ and $h(c)$ to give $\tilde{g}(p)$ and $\tilde{h}(c)$.
- We include the drift parameter in the model.
- The model is essentially

$$\begin{aligned}\log(\text{rate}) = \text{Age on rate scale} &+ \text{Linear drift} \\ &+ \text{non-linear Period Effects} \\ &+ \text{non-linear Cohort effects}\end{aligned}$$

Other software for spline based APC models

apcfit Model incidence, difficult to extrapolate.

apcspline Model incidence, specific types of models.

- Neither of these commands allow for easy combination with survival models to project future prevalence.

The apcmodel commands

The command is still in development

apcsetup	Set up data for modelling incidence.
apcmodel	Model incidence (AP, AC, APC models).
predict	Predict rates, number of cases etc.
apcplot	Various plots (oberved and model based rates)
apcprevalence	Combine incidence with survival to project future prevalence.

The apcmodel commands

The command is still in development

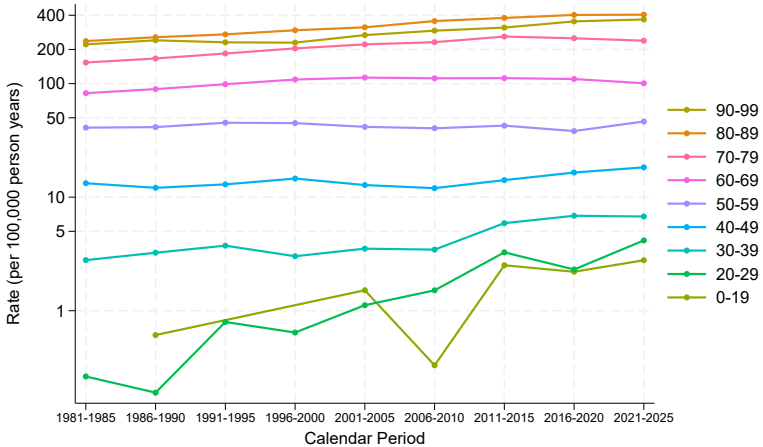
apcsetup	Set up data for modelling incidence.
apcmodel	Model incidence (AP, AC, APC models).
predict	Predict rates, number of cases etc.
apcplot	Various plots (oberved and model based rates)
apcprevalence	Combine incidence with survival to project future prevalence.

- The apcmodel commands requires.

gensplines	Create spline basis functions for age, period and cohort.
partpred	Partial predictions.
stpm3	Flexible parametric survival models (for prevalence projections).

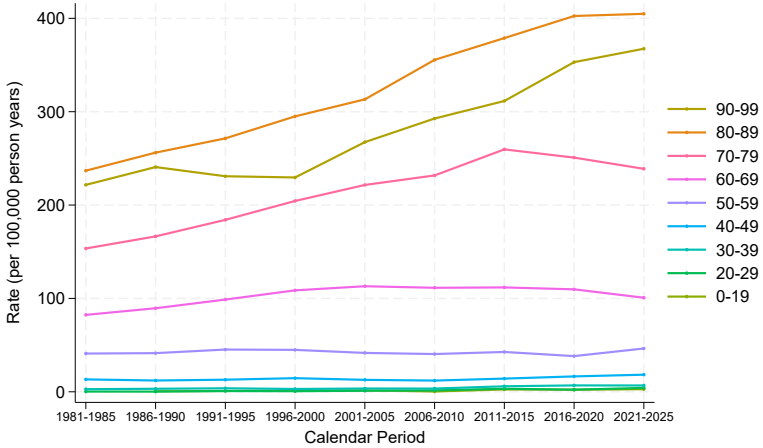
Example: Female Colon Cancer: Period-Age Plot

```
// descriptive pa plot: log scale
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) legendreverse ///
            obslineopts(lpattern(solid) lwidth(*1.5))                  ///
            xlabel(1(1)9,valuelabel labsize(*0.75))
```



Example: Female Colon Cancer: Period-Age Plot

```
// descriptive pa plot: untransformed scale
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) legendreverse nolog ///
            obslineopts(lpattern(solid) lwidth(*1.5))                        ///
            xlabel(1(1)9,valuelabel labsize(*0.75))
```



Modelling incidence (without projections)

```
// age-period model
apcmodel D, model(AP) age(A) period(P) personyears(Y)   ///
            dfperiod(6) dfage(6)

// age-cohort model
apcmodel D, model(AC) age(A) cohort(C) personyears(Y)   ///
            dfcohort(6) dfage(6)

// age-period-cohort model
// drift added to cohort effect
apcmodel D,  model(APC) age(A) period(P) cohort(C) personyears(Y) ///
             drift(cohort) dfcohort(6) dfage(6) dfperiod(6)


// age-period-cohort model
// drift added to period effect
apcmodel D,  model(APC) age(A) period(P) cohort(C) personyears(Y) ///
             drift(period) dfcohort(6) dfage(6) dfperiod(6)
```

Output for APC model

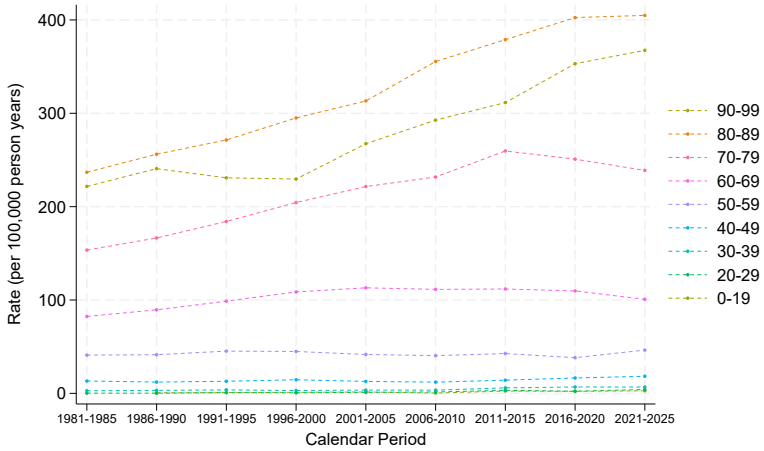
```
. apcmodel D,  age(A) period(P) cohort(C) personyears(Y)   ///
>              dfcohort(6) dfage(6) dfperiod(6) ///
>              model(APC) drift(cohort)
```

(*output omitted*)

D	Coefficient	OIM std. err.	z	P>|z|	[95% conf.	interval]
__A0	-5.764339	.0452728	-127.32	0.000	-5.853072	-5.675606
__A1	-20.8166	.4997185	-41.66	0.000	-21.79603	-19.83717
__A2	2.2801	.2555561	8.92	0.000	1.779219	2.780981
__A3	-2.995109	.0589273	-50.83	0.000	-3.110604	-2.879613
__A4	-1.486428	.0562061	-26.45	0.000	-1.59659	-1.376266
__A5	-.3688832	.0398453	-9.26	0.000	-.4469785	-.2907879
__A6	.8833219	.0794227	11.12	0.000	.7276562	1.038988
__C1	21.64966	2.548489	8.50	0.000	16.65472	26.64461
__C2	-3.033786	.3567771	-8.50	0.000	-3.733056	-2.334516
__C3	3.796299	.4504986	8.43	0.000	2.913338	4.67926
__C4	2.810328	.3290049	8.54	0.000	2.16549	3.455166
__C5	1.35478	.2372487	5.71	0.000	.8897813	1.819779
__Cdrift	.0101809	.000376	27.07	0.000	.0094439	.010918
__P1	-1.686107	.5600258	-3.01	0.003	-2.783738	-.5884766
__P2	.0863159	.0767026	1.13	0.260	-.0640185	.2366504
__P3	-.3202889	.1250678	-2.56	0.010	-.5654174	-.0751604
__P4	-.2546993	.0809035	-3.15	0.002	-.4132671	-.0961314
__P5	-.1502352	.067071	-2.24	0.025	-.281692	-.0187784
ln(Y)	1	(exposure)				

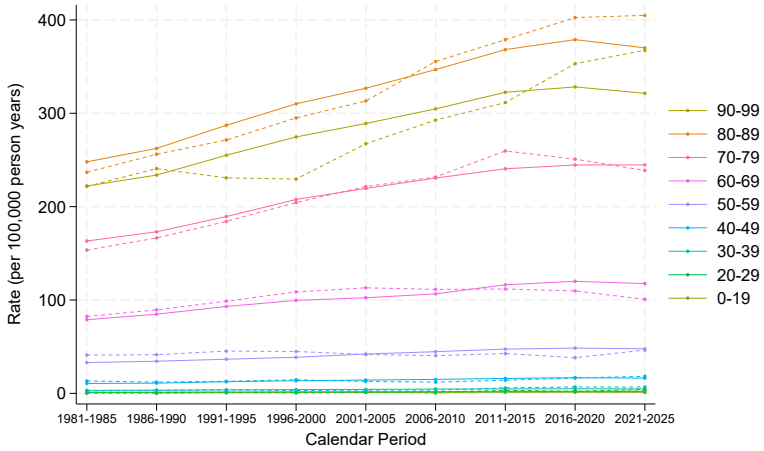
Fit from Age-Period Model

```
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) ///
            legendreverse nolog predicted
```



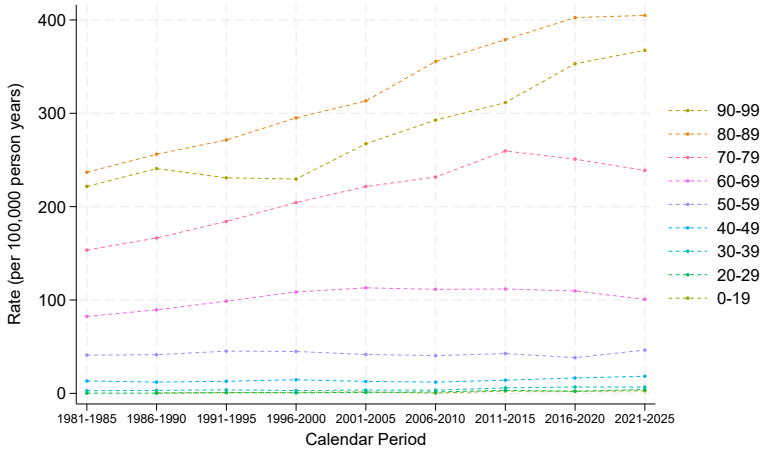
Fit from Age-Period Model

```
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) ///
            legendreverse nolog predicted
```



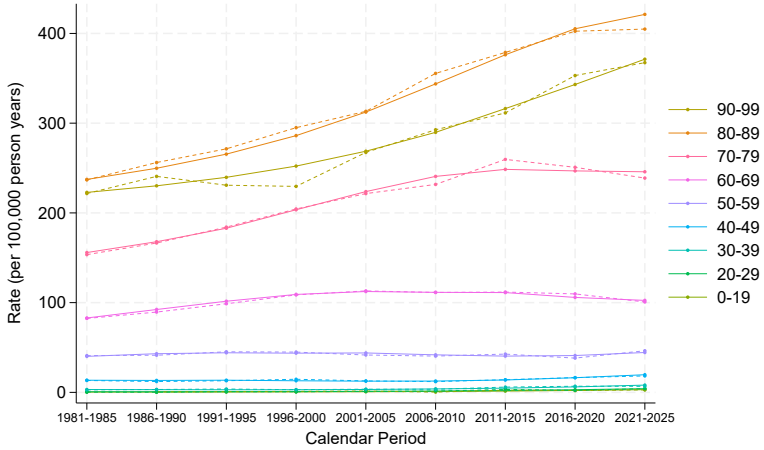
Fit from Age-Cohort Model

```
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) ///
            legendreverse nolog predicted
```



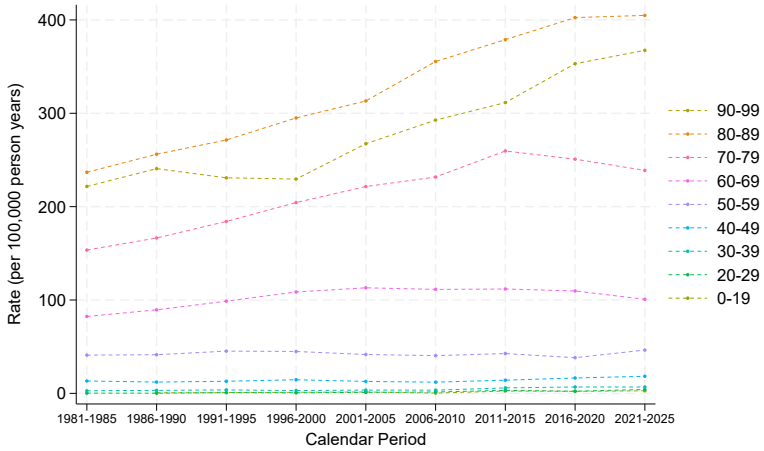
Fit from Age-Cohort Model

```
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) ///
            legendreverse nolog predicted
```



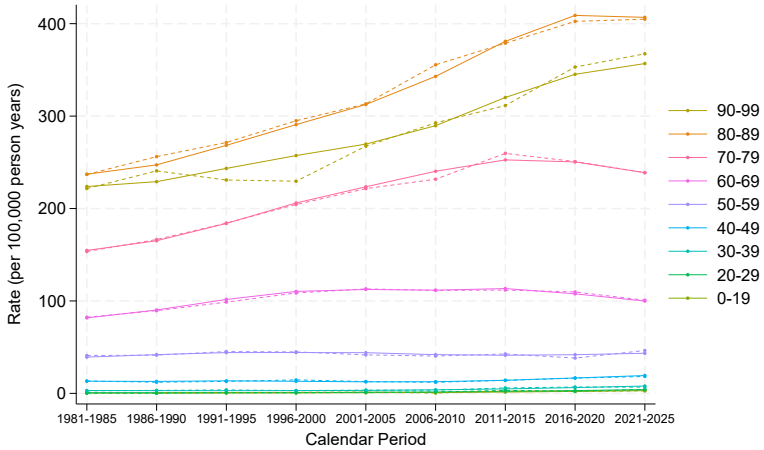
Fit from Age-Period-Cohort Model

```
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) ///
            legendreverse nolog predicted
```



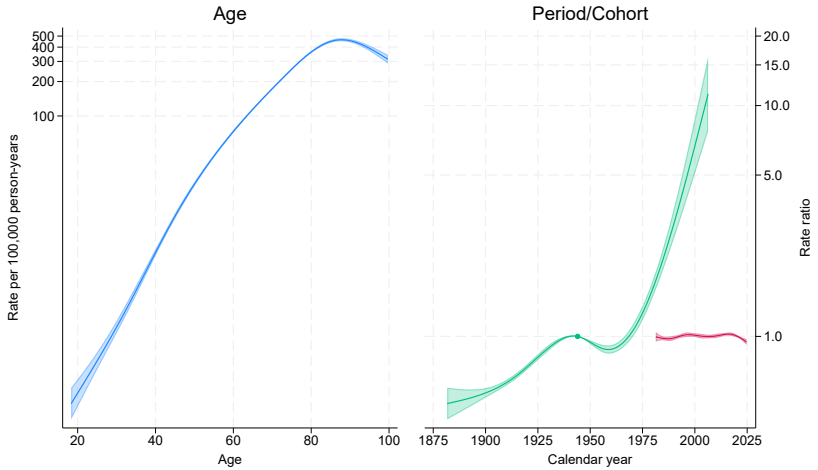
Fit from Age-Period-Cohort Model

```
apcplot pa, agegroup(0 20(10)100) yeargroup(periodgroup) ///
            legendreverse nolog predicted
```



Summary fit from APC model

```
apcplot apc, ratioxlabel(1875(25)2025)
```



Projecting future incidence

- Different modelling assumptions:
  - Extrapolate linear trend (drift) with a log link function.
  - Extrapolate linear trend (drift) with a power 1/5 link function[6].
  - Extrapolate linear trend (log or power 1/5) with dampening[3].
  - Moving last boundary knot in x years to estimate a more local linear trend. Either the log or power 1/5 link can be used[7].
  - Assume no change in future incidence rates.

Projecting future incidence

- Different modelling assumptions:
  - Extrapolate linear trend (drift) with a log link function.
  - Extrapolate linear trend (drift) with a power 1/5 link function[6].
  - Extrapolate linear trend (log or power 1/5) with dampening[3].
  - Moving last boundary knot in x years to estimate a more local linear trend. Either the log or power 1/5 link can be used[7].
  - Assume no change in future incidence rates.
- All the above can be implemented in `apcmodel` using different combination of the the `link()`, extrapmethod(), lastknot() and `dampening()` options.

Projecting future incidence: Code

Extrapolate drift with power 1/5 link

```
apcmodel D, age(A) period(P) cohort(C) personyears(Y)  ///
            dfperiod(6)  dfcohort(6) dfage(6)          ///
            drift(period) extrapfrom(2026)             ///
            link(power5) extrapmethod(drift)
```

Move last knot in 10 years and extrapolate recent trend

```
apcmodel D, age(A) period(P) cohort(C) personyears(Y) link(power5) ///
            dfperiod(6)  dfcohort(6) dfage(6)                      ///
            drift(period) extrapfrom(2026)                         ///
            lastknot(2016) extrapmethod(ns)
```

Assume rates are as estimated for last follow-up year (constant)

```
apcmodel D, age(A) period(P) cohort(C) personyears(Y) link(log)    ///
            dfperiod(6)  dfcohort(6) dfage(6)                      ///
            drift(period) extrapfrom(2026)                         ///
            extrapmethod(constant)
```

Projecting future incidence: Code

Extrapolate drift with power 1/5 link

```
apcmodel D, age(A) period(P) cohort(C) personyears(Y)  ///
            dfperiod(6)  dfcohort(6) dfage(6)          ///
            drift(period) extrapfrom(2026)             ///
            link(power5) extrapmethod(drift)
```

Move last knot in 10 years and extrapolate recent trend

```
apcmodel D, age(A) period(P) cohort(C) personyears(Y) link(power5) ///
            dfperiod(6)  dfcohort(6) dfage(6)                      ///
            drift(period) extrapfrom(2026)                         ///
            lastknot(2016) extrapmethod(ns)
```

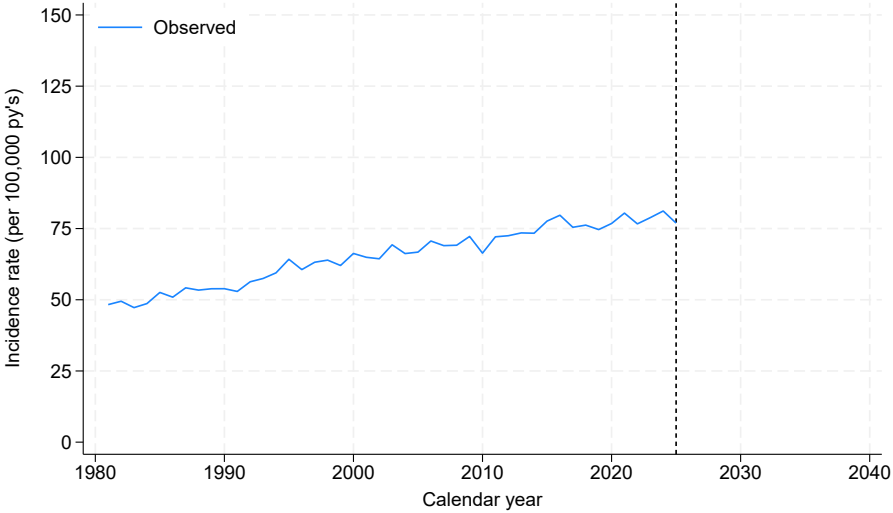
Assume rates are as estimated for last follow-up year (constant)

```
apcmodel D, age(A) period(P) cohort(C) personyears(Y) link(log)    ///
            dfperiod(6)  dfcohort(6) dfage(6)                      ///
            drift(period) extrapfrom(2026)                         ///
            extrapmethod(constant)
```

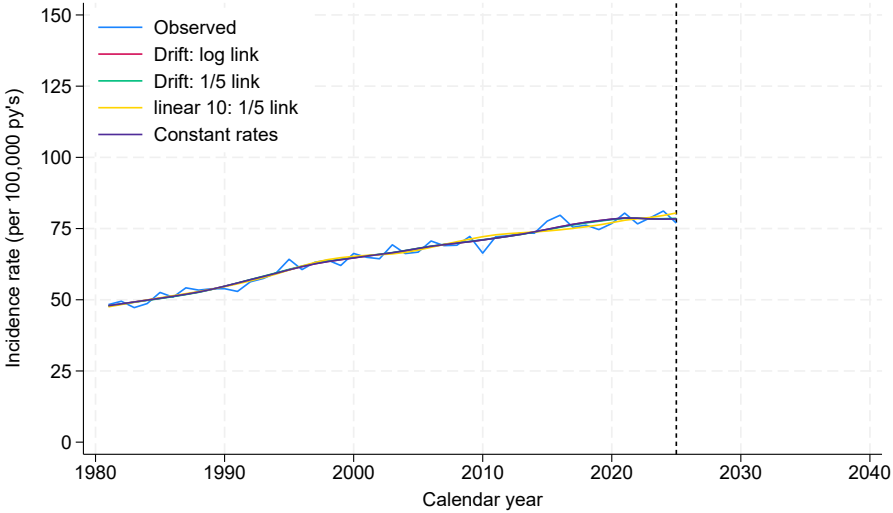
After fitting a model, predict rates and number of cases

```
predict rate_pow5,  rate per(100000)
predict cases_pow5, cases
```

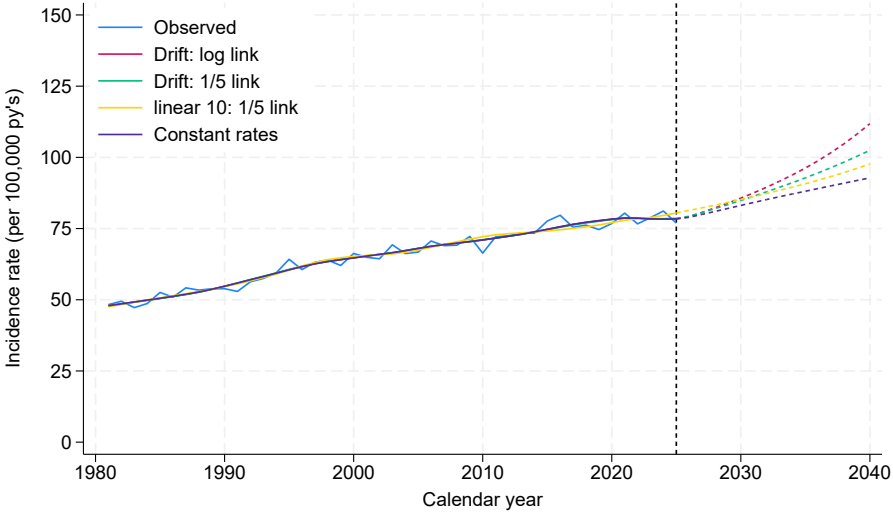
Projecting future incidence: rates



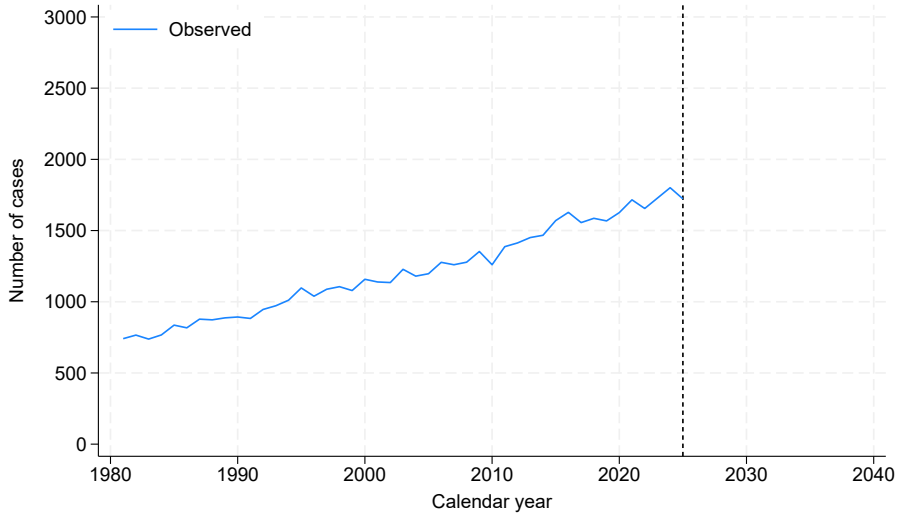
Projecting future incidence: rates



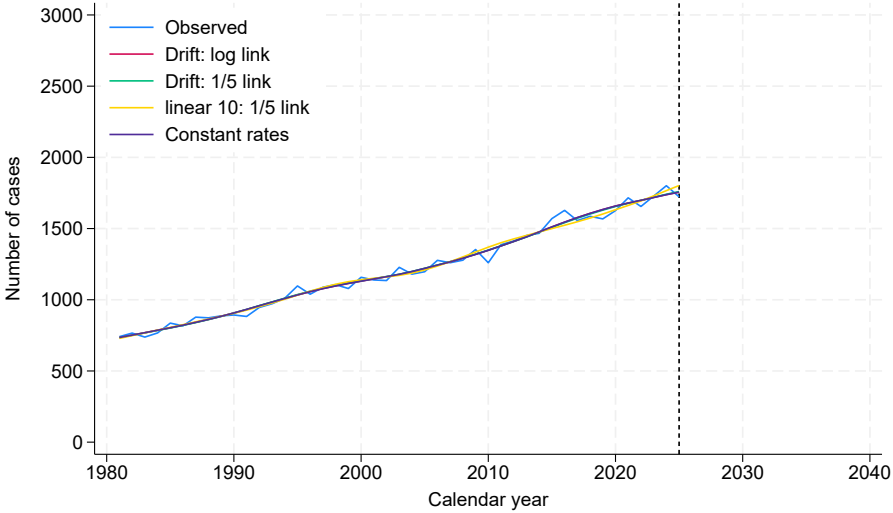
Projecting future incidence: rates



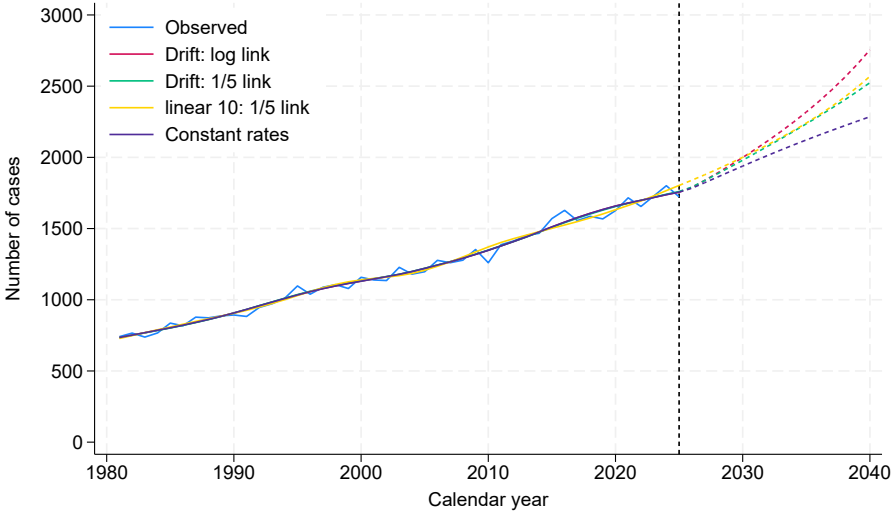
Projecting future incidence: cases



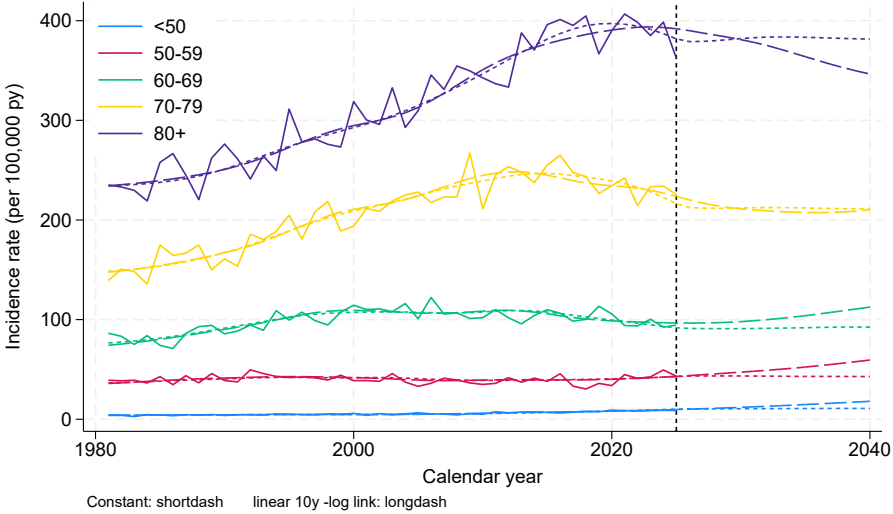
Projecting future incidence: cases



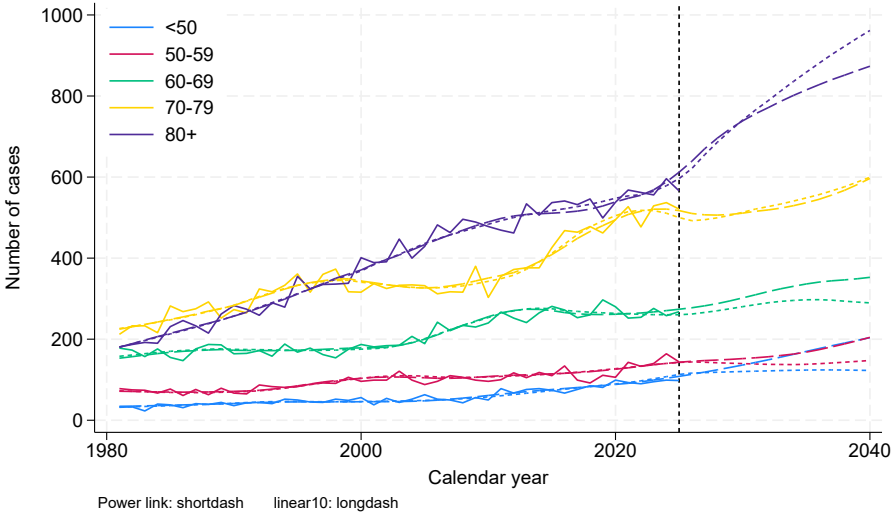
Projecting future incidence: cases



Projecting future incidence by age: rates/cases



Projecting future incidence by age: rates/cases



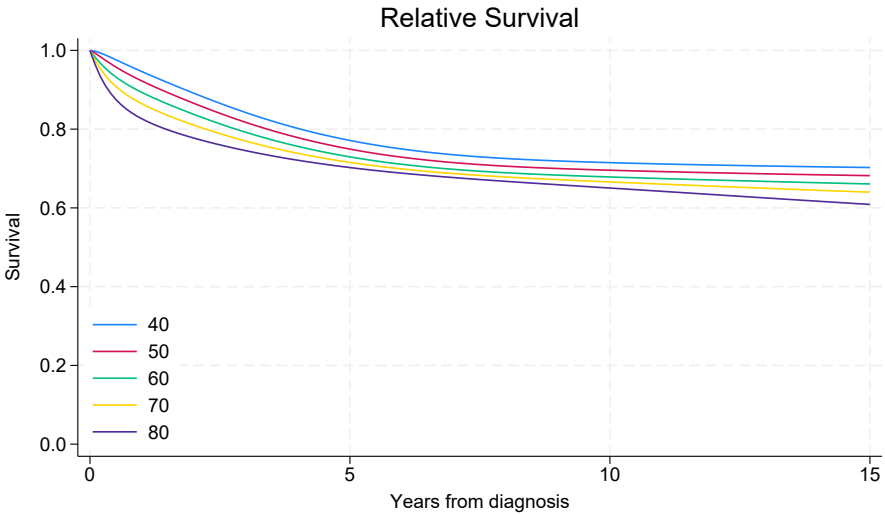
Survival model

- We use a flexible parametric survival models implemented in `stpm3`.
- As we want the extrapolate we make this an excess mortality model.
- The effect of time since diagnosis is modelled using natural splines.
- The effect of age is modelled using natural splines (with winsorizing).
- The proportional hazards assumption is relaxed through incorporation of time-dependent effects.
- In the example below we do not extrapolate future improvments in survival.
- However, we use a period analysis approach to obtain up-to-date estimates of survival[8].

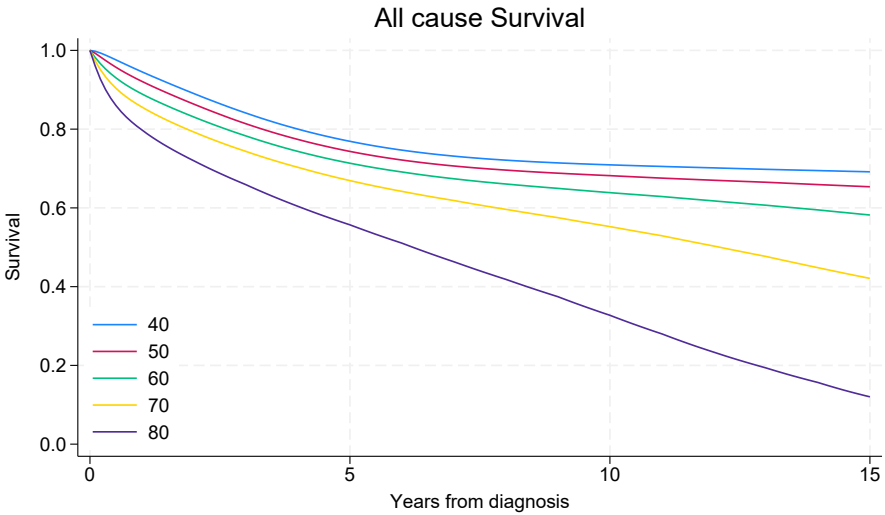
Survival predictions by age

```
// flexible parametric model
// fitted to "period analysis" data for up-to-date estimates
stpm3 @ns(agediag_nonint,df(3) winsor(2 98)),      /// natural splines for age
      scale(lncumhazard)                           /// log cumulative hazard scale
      df(5)                                        /// splines with 5df for follow-up time
      bhazard(rate)                                /// expected rates (relative survival model)
      tvc(@ns(agediag_nonint, df(3) winsor(2 98))) /// time-dependent effect of age
      dftvc(3)                                     //  3df for time-dependent effects
// store model, so can pass to apcprevalence command.
estimates store colon_female
```

Survival predictions by age



Survival predictions by age



Combining incidence and survival models

- Let $\widehat{I}(a, y)$ be the predicted incidence rate for age a in calendar year y and $Y(a, y)$ be the (predicted) population count for age a in calendar year y.
- The number of new cases for age a in calendar year y can be estimated as:

$$\widehat{D}(a, y) = \widehat{I}(a, y) \cdot Y(a, y)$$

- Let $\widehat{S}(t|a, y)$ be the predicted all cause survival probability t years after diagnosis for those diagnosed at age a in calendar year y.
- The number of prevalent cases at current age $a + t$ diagnosed at age a in calendar year y that are still alive in calendar year $y + t$ can be estimated:

$$P(a + t, y + t) = \widehat{D}(a, y) \cdot \widehat{S}(t|a, y)$$

- Note that $S(t|a, y)$ is an all cause survival probability. However, as the model needs to extrapolate into the future, a relative survival approach is useful, as it has been shown to lead to more accurate extrapolations[9].

Combining the counting method with modelling I

- We obtain (count) the number of prevalent cases at the last point of follow-up.
- At future year y^* the prevalence for age a^* will consist of
  - Individuals diagnosed before the end of follow-up and still alive
  - Individuals diagnosed between the end of follow-up and y^* and still alive.
- Let $\widetilde{D}(a, y, y_{end})$ denote the number of cases diagnosed at age a in calendar year y that are still alive at the end of follow-up in calendar year y_{end}. Then the number of prevalent cases at the end of future calendar year $y*$ can be calculated as:

$$\widehat{P}(a*, y*|a, y) = \widetilde{D}(a, y, y_{end}) \cdot \frac{\widehat{S}(y^* - y|a, y)}{\widehat{S}(y_{end} - y|a, y)}$$

- This is a conditional survival probability, i.e. the probability of surviving to year, $y*$ *given* diagnosis at year y at age a, and still alive at year y_{end}.

Combining the counting method with modelling II

- For those diagnosed after year y_{end} at year y aged a, the number of new cases is....

$$\widehat{P}(a*, y*|a, y) = \widehat{I}(a, y) \cdot \widehat{Y}(a, y) \cdot \widehat{S}(y^* - y|a, y) \quad (1)$$

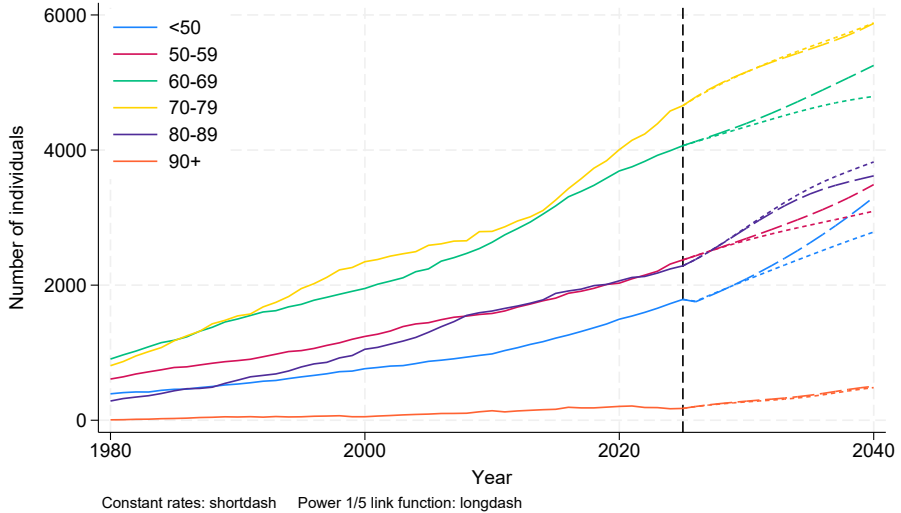
- We can then sum over age and/or year groups etc to obtain prevalence estimates.

Examples: apcprevalence

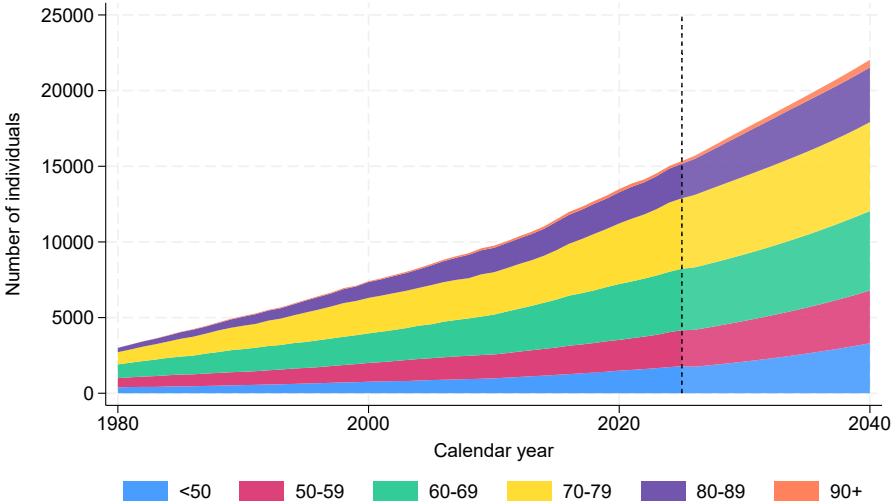
```
use results/statatalk/predicted_incidence_cases, clear
apcprevalence, cases(cases_pow5 cases_pow5_lin10 cases_cons) /// predictions from incidence models
               survmodel(surv_model)                         /// survival model
               pmmaxyear(2025)                               /// maximum year in expected rates
               extrapfrom(2026)                              /// extrap from year
               extrapto(2040)                                /// extrap to year
               pmother(sex)                                  /// strata in lifetable
               agemodel(agediag_nonint)                      /// age variable in model
               popmort(${lifetable})                         /// popmort file
               duration(1 5 10)                              /// additional x-year prevalence
               saving(`predictedprev', replace)              //  results file

// collapse for prediction over specific strata
// e.g. by agegroup and calendar year
frame create by_agegroup_year
frame by_agegroup_year {
  use `predictedprev', clear
  rename A age
  rename P year
  recode age (min/49=1) (50/59=2) (60/69=3) (70/79=4) (80/89=5) (90/max=6), gen(agegroup)
  collapse (sum) N*, by(year agegroup)
}
```

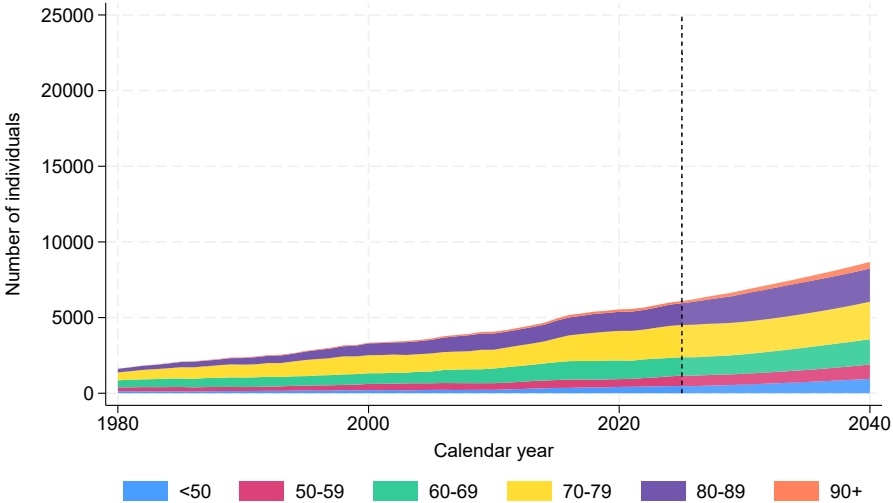
Examples: Future prevalence by agegroup



Examples: Future prevalence by agegroup



Examples: Future prevalence by agegroup (within 5 years of diagnosis)



Summary

- Projections of incidence and prevalence important for future planning.
- The `apcmodel` set of commands makes projection of future incidence and prevalence relatively easy.
- Enables comparison of different scenarios and assumptions.
- Command still in development.

References I

A. Verdecchia, G. D. Angelis, and R. Capocaccia, "Estimation and projections of cancer prevalence from cancer registry data," *Statistics in Medicine*, vol. 21, p. 3511–3526, Oct. 2002.

B. Carstensen, "Age–period–cohort models for the lexis diagram," *Statistics in Medicine*, vol. 26, p. 3018–3045, Dec. 2006.

P. D. Sasieni, "Age–period–cohort models in stata," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 12, p. 45–60, Mar. 2012.

M. J. Rutherford, P. C. Lambert, and J. R. Thompson, "Age–period–cohort modeling," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 10, p. 606–627, Dec. 2010.

D. Clayton and E. Schifflers, "Models for temporal variation in cancer rates. ii: Age–period–cohort models," *Statistics in Medicine*, vol. 6, no. 4, p. 469–481, 1987.

References II


B. Møller, H. Fekjær, T. Hakulinen, H. Sigvaldason, H. H. Storm, M. Talbäck, and T. Haldorsen, "Prediction of cancer incidence in the nordic countries: empirical comparison of different approaches," *Statistics in Medicine*, vol. 22, p. 2751–2766, Aug. 2003.


M. J. Rutherford, J. R. Thompson, and P. C. Lambert, "Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines," *The International Journal of Biostatistics*, vol. 8, Nov. 2012.


H. Brenner and T. Hakulinen, "Up-to-date long-term survival curves of patients with cancer by period analysis," *Journal of Clinical Oncology*, vol. 20, p. 826–832, Feb. 2002.


T. M.-L. Andersson, M. J. Rutherford, and P. C. Lambert, "Illustration of different modelling assumptions for estimation of loss in expectation of life due to cancer," *BMC Medical Research Methodology*, vol. 19, no. 1, 2019.