A practical approach to fitting cancer survival models when data can't move across borders

Paul C Lambert[1,2], Mark J Rutherford[3], Tor Åge Myklebust[1,4]

[1]Cancer Registry of Norway, FHI, Norway
[2]Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[3]Biostatistics Research Group, Population Health Sciences, University of Leicester, UK
[4]Dept. of Research and Innovation, Møre and Romsdal Hospital Trust, Ålesund, Norway

ANCR symposium 2024, Bodø, Norway, 30th August 2024

Slides: pclambert.net/pdf/Paul_Lambert_ANCR2024.pdf
Example: pclambert.net/software/standsurv/models_different_countries

Cancer Registry of Norway Karolinska Institutet

Introduction

- It is getting harder to share data between countries, making international comparisons more difficult.
- Here, I focus on survival analysis.
  - Generally need individual level data
  - Sometimes we need/want to use statistical modelling approaches (e.g. recent NORDCAN Survival Studies).
- NORDCAN.R showed how a federated approach could be applied.
  - Data analysed separately in each country
  - Aggregated/summary data sent to IARC
- Here I will explore something similar for a modelling approach.

Single model or separate models?

- We have choices.
  1. Fit a separate model for each country.
  2. Fit a single joint model to all countries.
- A single model can be more efficent as we can 'borrow strength' between countries.
  - However, it requires data to be in one place or to use a full federated learning approach.
- If we have large data then we are happier to fit separate models.
- A joint model with interactions between country and all covariates (and time) is equivalent to separate models.

Options for a modelling approach

1. Fit model separately in each country
   - Extract statistics of interest (e.g. 5 year relative survival)
   - Send to hub

Options for a modelling approach

1. Fit model separately in each country
   - Extract statistics of interest (e.g. 5 year relative survival)
   - Send to hub
2. Fit model separately in each country
   - Save model object (containing model parameters etc)
   - Send to hub

Options for a modelling approach

1. Fit model separately in each country
   - Extract statistics of interest (e.g. 5 year relative survival)
   - Send to hub
2. Fit model separately in each country
   - Save model object (containing model parameters etc)
   - Send to hub
3. Federated learning
   - Hub defines model
   - Parameters sent to each node
   - Aggregated model information sent back
   - Parameters updated
   - Repeat until convergence

Options for a modelling approach

1. Fit model separately in each country
   - Extract statistics of interest (e.g. 5 year relative survival)
   - Send to hub
2. Fit model separately in each country
   - Save model object (containing model parameters etc)
   - Send to hub
3. Federated learning
   - Hub defines model
   - Parameters sent to each node
   - Aggregated model information sent back
   - Parameters updated
   - Repeat until convergence

We should only choose (3) if we need to. Often (2) will be sufficient.

Options for a modelling approach

2. Fit model separately in each country
   - Save model object (containing model parameters etc)
   - Send to hub

Example

- Uses entirely simulated (synthetic) data, so code and data available for people to try for themselves.
- Comparing Country A and Country B.
- I assume I do not have access to data in Country A.
- Detailed example on my webpage

Model fitted in Country A

- I have a colleague willing to run code in Country A

Fit model in Country A

```
// Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>                 bhazard(rate) tvc(i.sex @ns(age,df(3)))
// Save model object
. estimates save countryA.ster
```

Model fitted in Country A

- I have a colleague willing to run code in Country A

Fit model in Country A

```
 // Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>                 bhazard(rate) tvc(i.sex @ns(age,df(3)))
// Save model object
. estimates save countryA.ster
```

Model fitted in Country A

- I have a colleague willing to run code in Country A

Fit model in Country A

```
 // Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>                 bhazard(rate) tvc(i.sex @ns(age,df(3)))
// Save model object
. estimates save countryA.ster
```

- My colleague sends this file to me in Country B (or elsewhere)

What's stored in .ster file?

- The ingredients needed to predict survival etc from the model.
  - Names of covariates included in the model
  - Parameter estimates and variances
  - Knot locations for spline functions.
  - Various other details (Number of parameters, sample size, likelihood etc)

What's stored in .ster file?

- The ingredients needed to predict survival etc from the model.
  - Names of covariates included in the model
  - Parameter estimates and variances
  - Knot locations for spline functions.
  - Various other details (Number of parameters, sample size, likelihood etc)

Crucially it contains no individual level data

Analyse data in Country B

```
// Load data
. use https://www.pclambert.net/data/CountryB, clear

// Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>               bhazard(rate) tvc(i.sex @ns(age,df(3)))

// standardized relative survival for Country B
. standsurv RS_B, surv timevar(tt) ci frame(RS, replace)

// Load model object for Country A (BUT NOT DATA)
. estimate use CountryA

// standardized relative survival for Country A
// NOTE: standardized to age/sex distribution of Country B
. standsurv RS_A, surv frame(RS, merge) ci
```

Analyse data in Country B

```
// Load data
. use https://www.pclambert.net/data/CountryB, clear

// Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>               bhazard(rate) tvc(i.sex @ns(age,df(3)))

// standardized relative survival for Country B
. standsurv RS_B, surv timevar(tt) ci frame(RS, replace)

// Load model object for Country A (BUT NOT DATA)
. estimate use CountryA

// standardized relative survival for Country A
// NOTE: standardized to age/sex distribution of Country B
. standsurv RS_A, surv frame(RS, merge) ci
```

Analyse data in Country B

```
// Load data
. use https://www.pclambert.net/data/CountryB, clear

// Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>                bhazard(rate) tvc(i.sex @ns(age,df(3)))

// standardized relative survival for Country B
. standsurv RS_B, surv timevar(tt) ci frame(RS, replace)

// Load model object for Country A (BUT NOT DATA)
. estimate use CountryA

// standardized relative survival for Country A
// NOTE: standardized to age/sex distribution of Country B
. standsurv RS_A, surv frame(RS, merge) ci
```

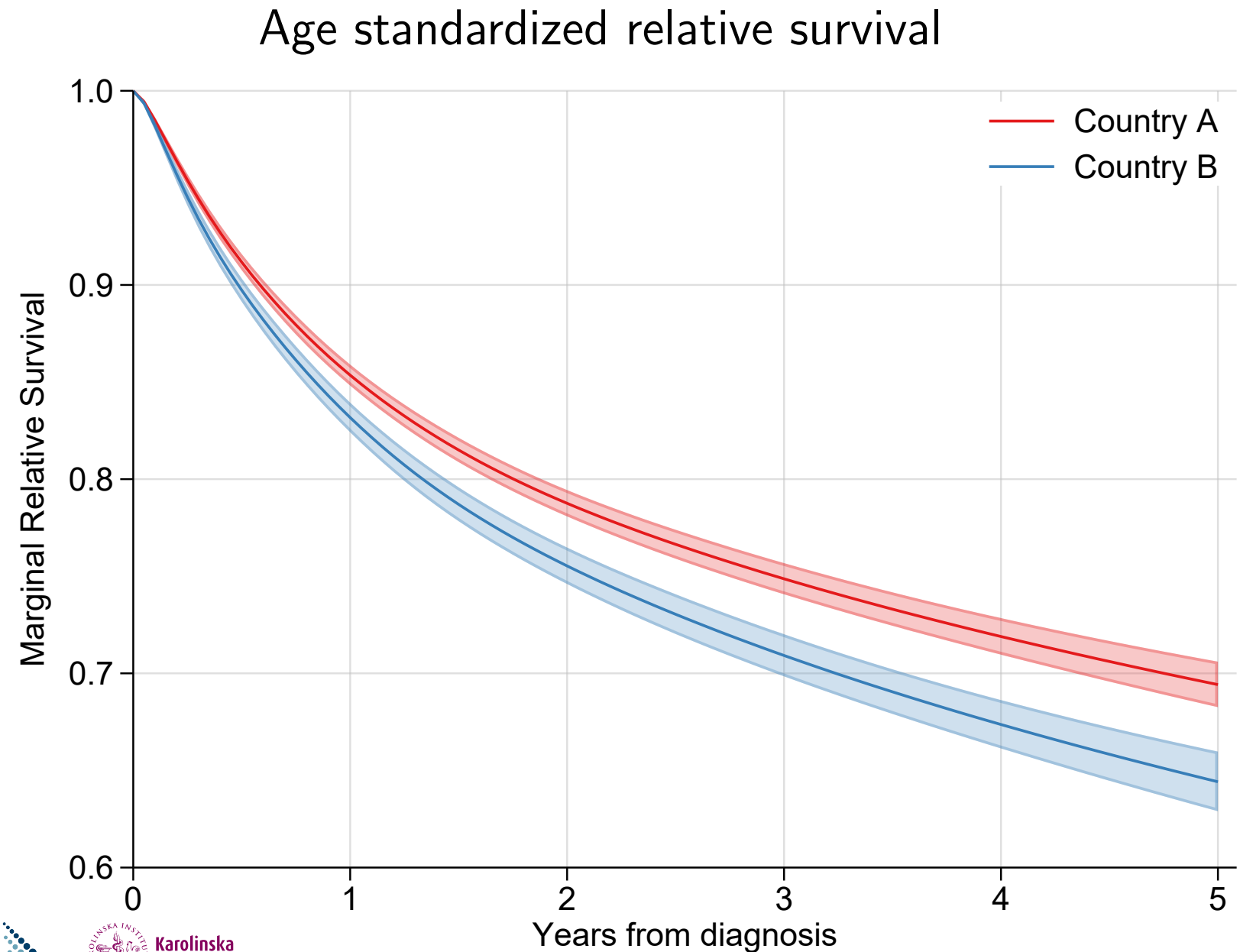
Analyse data in Country B

```
// Load data
. use https://www.pclambert.net/data/CountryB, clear

// Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>               bhazard(rate) tvc(i.sex @ns(age,df(3)))

// standardized relative survival for Country B
. standsurv RS_B, surv timevar(tt) ci frame(RS, replace)

// Load model object for Country A (BUT NOT DATA)
. estimate use CountryA

// standardized relative survival for Country A
// NOTE: standardized to age/sex distribution of Country B
. standsurv RS_A, surv frame(RS, merge) ci
```

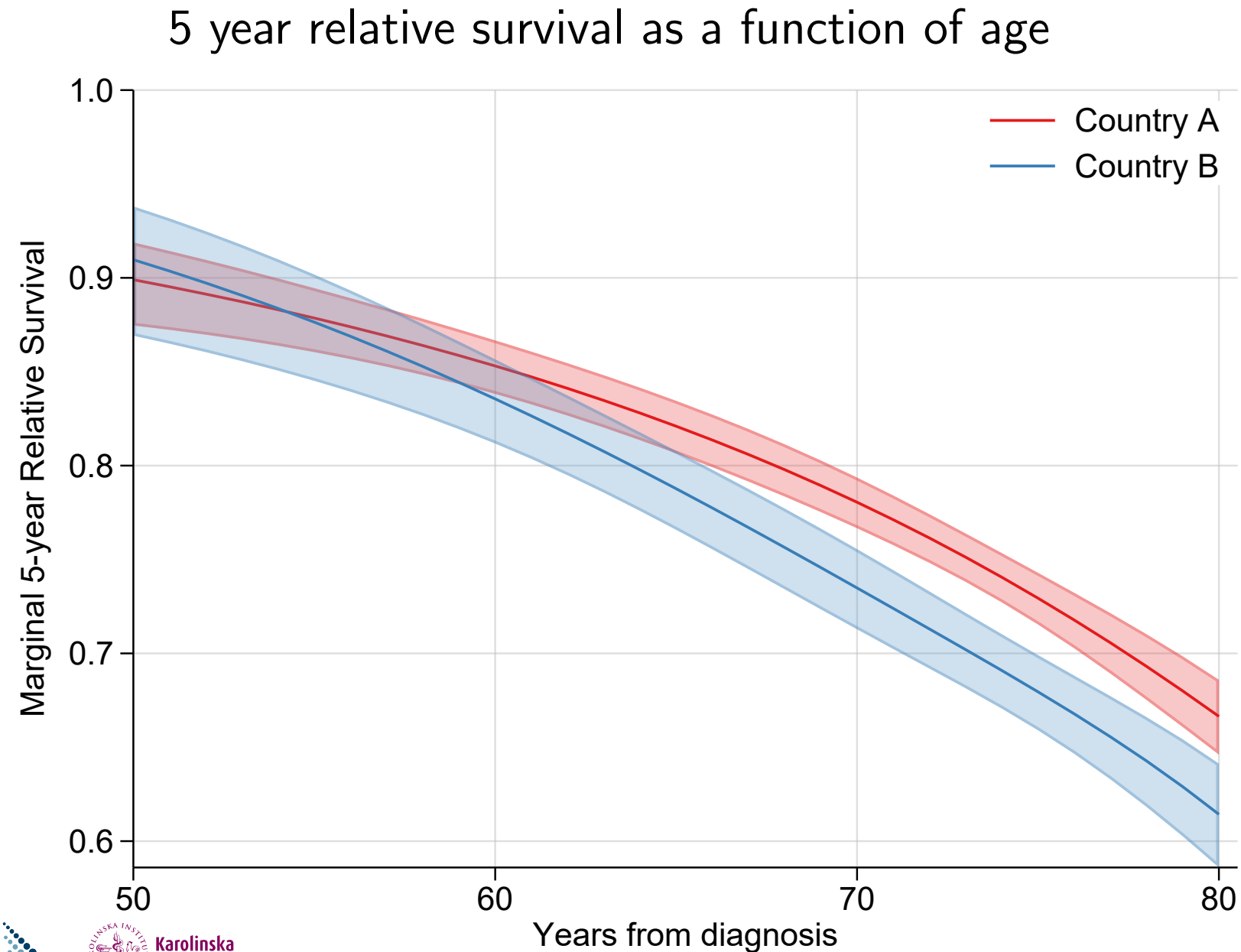
Analyse data in Country B

```
// Load data
. use https://www.pclambert.net/data/CountryB, clear

// Fit relative survival model
. stpm3 i.sex##@ns(age,df(3)), scale(lncumhazard) df(3) ///
>                bhazard(rate) tvc(i.sex @ns(age,df(3)))

// standardized relative survival for Country B
. standsurv RS_B, surv timevar(tt) ci frame(RS, replace)

// Load model object for Country A (BUT NOT DATA)
. estimate use CountryA

// standardized relative survival for Country A
// NOTE: standardized to age/sex distribution of Country B
. standsurv RS_A, surv frame(RS, merge) ci
```

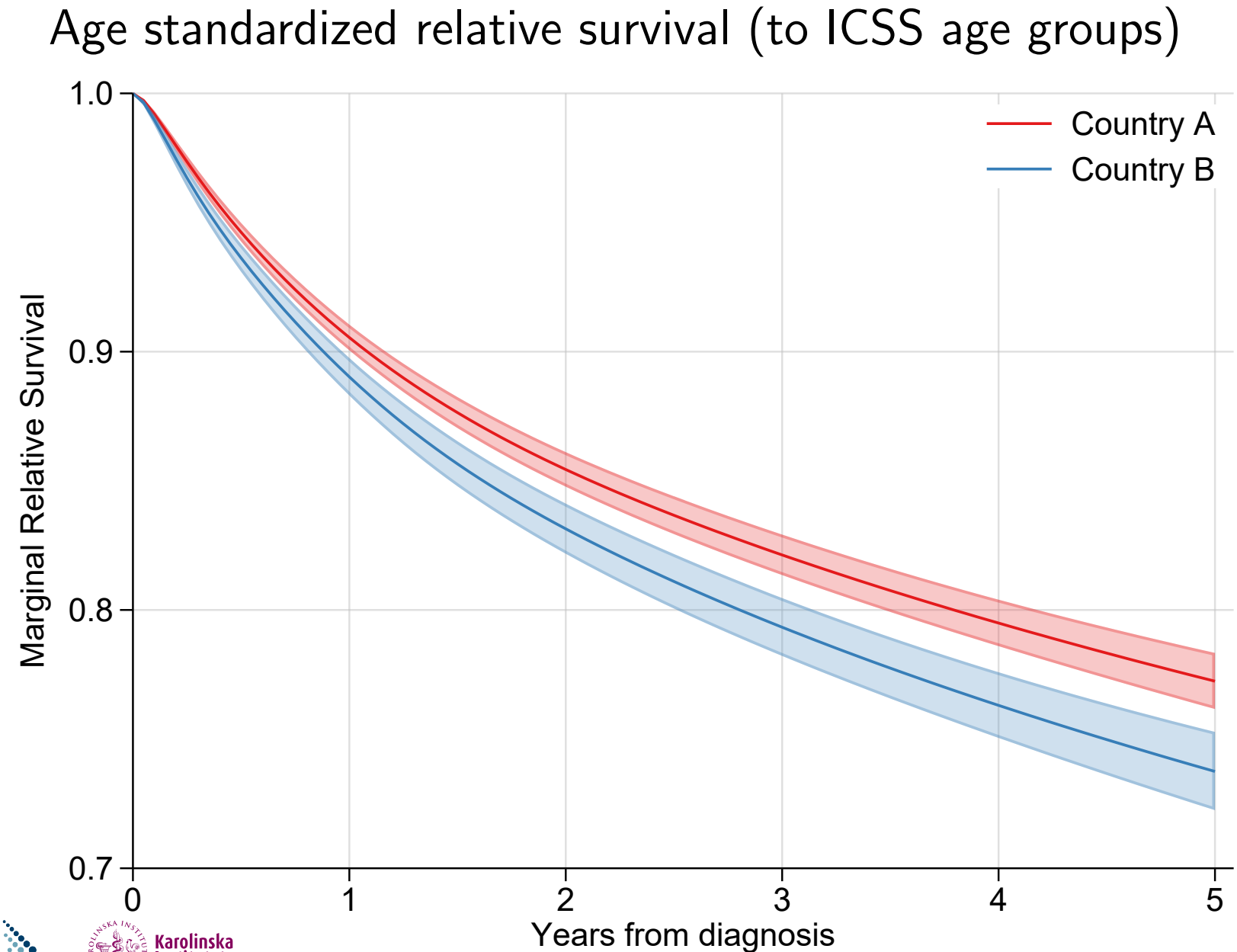
Results (of simulated data)



Results (of simulated data)



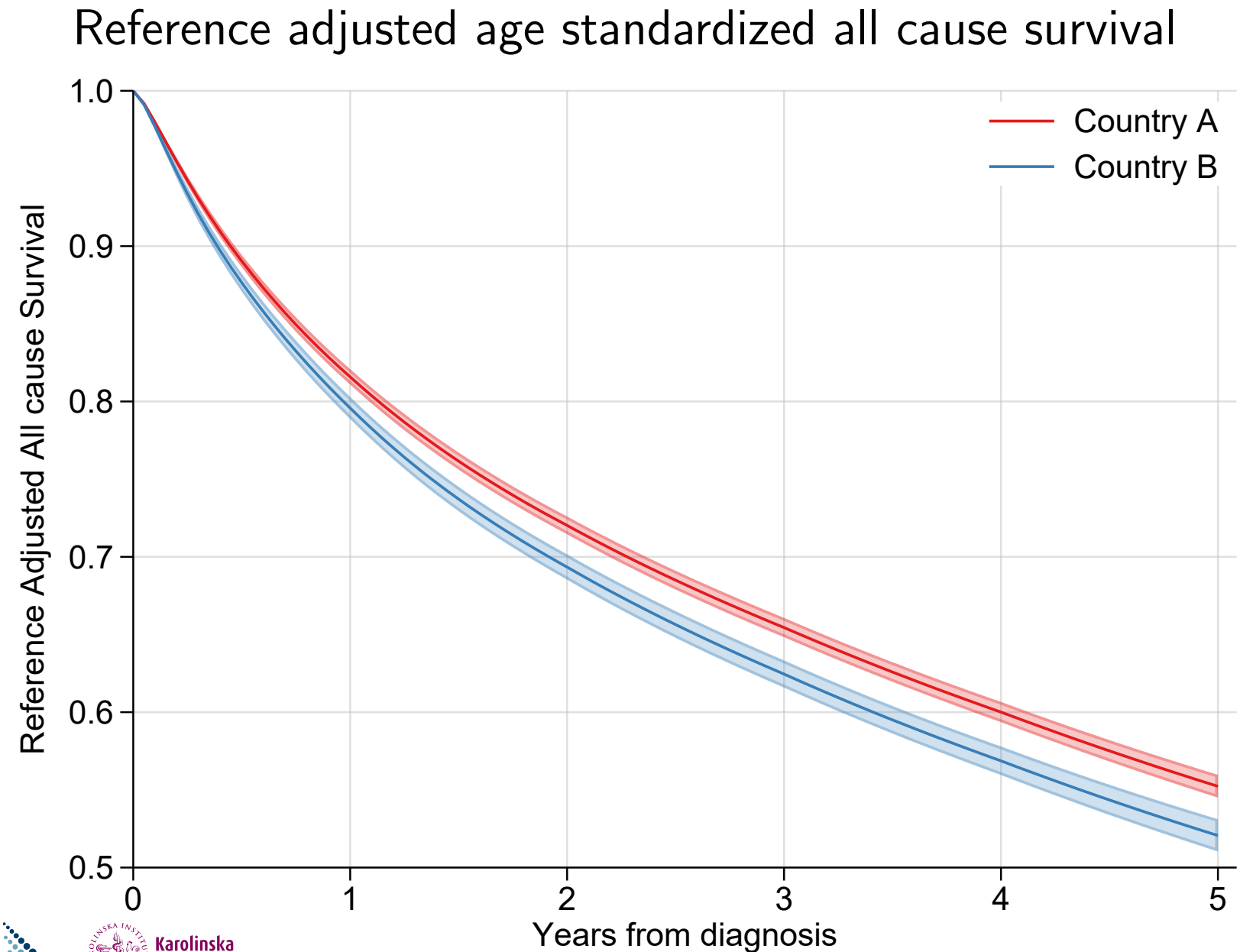
Results (of simulated data)



Cancer Registry of Norway

Karolinska Institutet

Results (of simulated data)



Standardization

- In the example standardization was to the age/sex distribution of Country B
- Easy to standardize to an external reference, e.g ICSS.
- Also possible to standardize to age/sex distribution of Country A with some summary (aggregated) information.

See extended example on my webpage

Discussion

- Simple way to fit separate models, but still obtain useful, and comparable, summaries from those models.
- More flexible than each country producing summaries and just sending those.
- Data quality, inclusion/exclusion criteria, consistency of variable naming/labelling very important.
- A full federated learning approach would give more control and ability to fit a combined model,
- However, this is a simple approach, that works.
- More details on my webpage.