When Notebooks are not Enough: Constructing Workflows for Reproducible Analytics

Andriy Koval

Matrix Institute Colloquium Series

University of Victoria

2018-10-31

github.com/andkov/ipdln-2018-hackathon

When notebooks are not enough

Last time at the Matrix Institute (2018-10-17)

- (Data) Science is about creating software!
- Tradeoff "Exploration vs Engineering"
- Limitations of Notebooks (by Neil Ernst)
  - Parameter configuration
  - Hidden state
  - Longevity and version control
  - Testing and modularity
  - Notebook carpentry

Today: Do reproducible projects overcome these limitations?

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff
0.1 Modeling form: univariate logistic regression with categorical predictors
0.2 Graphical form: faceted scatterplot in ggplot2
0.3 Coloring book: Mapping informed expectations from predictors onto color

B. Workflow Highlights

1.0 "Let no one ignorant of geometry enter": (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Autonomous phases: data cleaning, statistical modelling, graph production
1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
1.4 Two essential means of production: knitr::stitch() vs rmarkdown::render()

C. Conclusions

2.0 Different than Notebooks: sacrifices simplicity for agility via layers of isolation
2.1 R (+ .Rmd) = .html (+ .pdf) : moving away from *data playing* towards *data science*
2.2 Reproducible projects: moving away from notebooks towards software
2.3 Looking back to Neil Ernst talk:

- Parameters and configuration
- Hidden state
- Longevity and version control
- Testing and modularity
- Notebook carpentry

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff

International Population Data Linkage Conference 2018
The LIDIC Hackathon: LInked Data Innovation Challenge

Information for Participants

Date and Time: September 11, 2018 afternoon

Sponsors: We are grateful for sponsorship of this workshop by Statistics Canada and IBM.

Description: Participants will engage in a team-based analysis of a complex, linked, synthesized dataset provided by Statistics Canada. This synthesized data base links socioeconomic and mortality data representing the Canadian population. The data based was derived from existing linked data available at Statistics Canada.

Objectives:

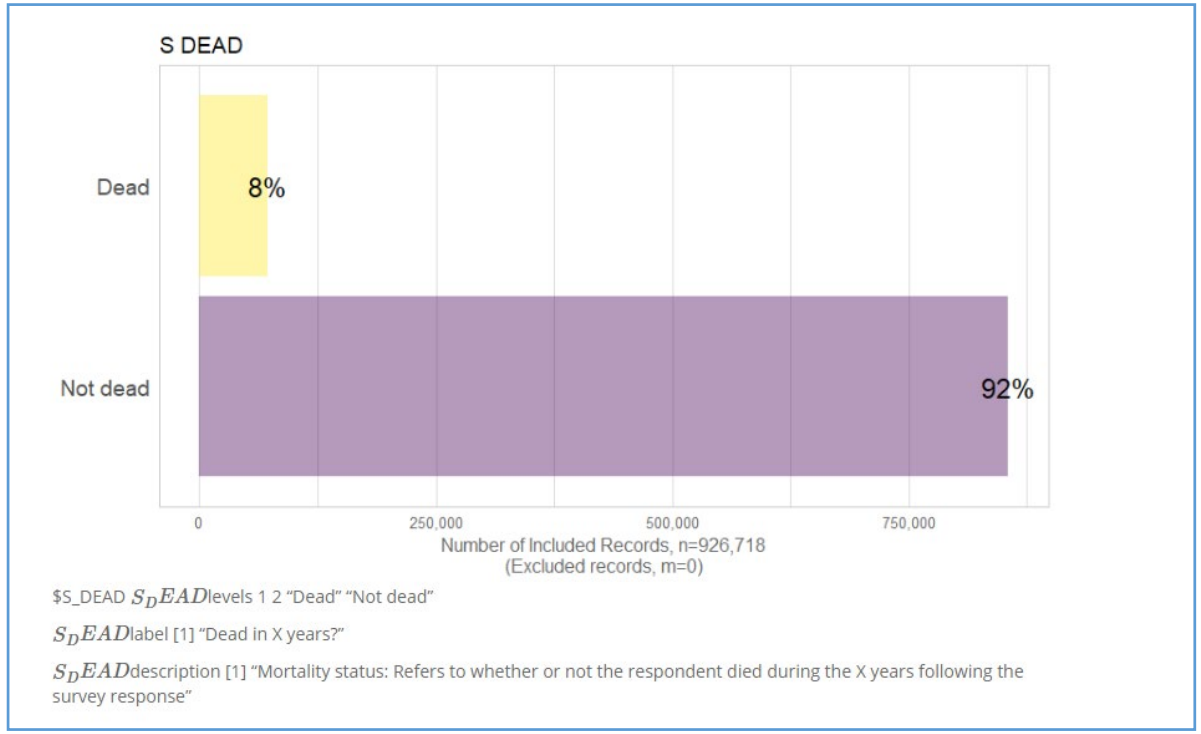
- To encourage innovative thinking about complex linked databases
- To stimulate interdisciplinary and inter-jurisdictional data collaborations
- To facilitate an environment for creative thinking about data
- To promote networking amongst participants

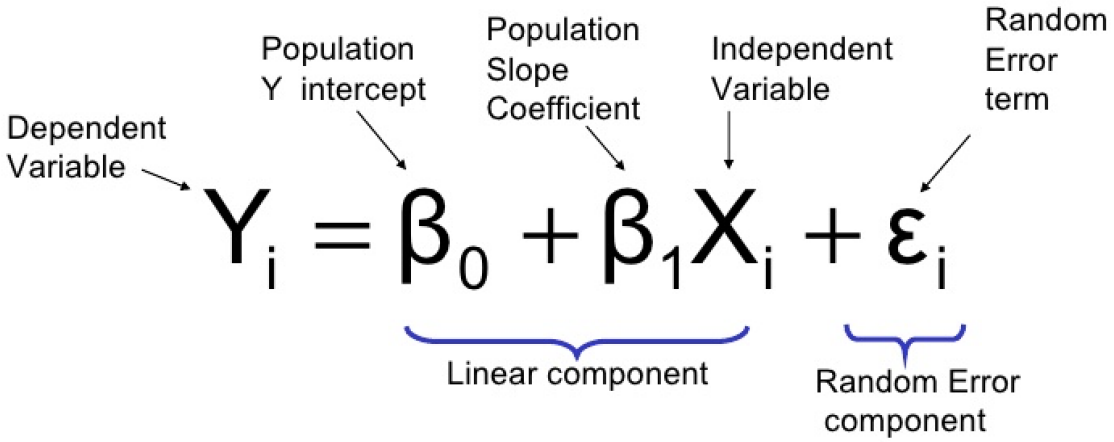
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Dead in X years





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

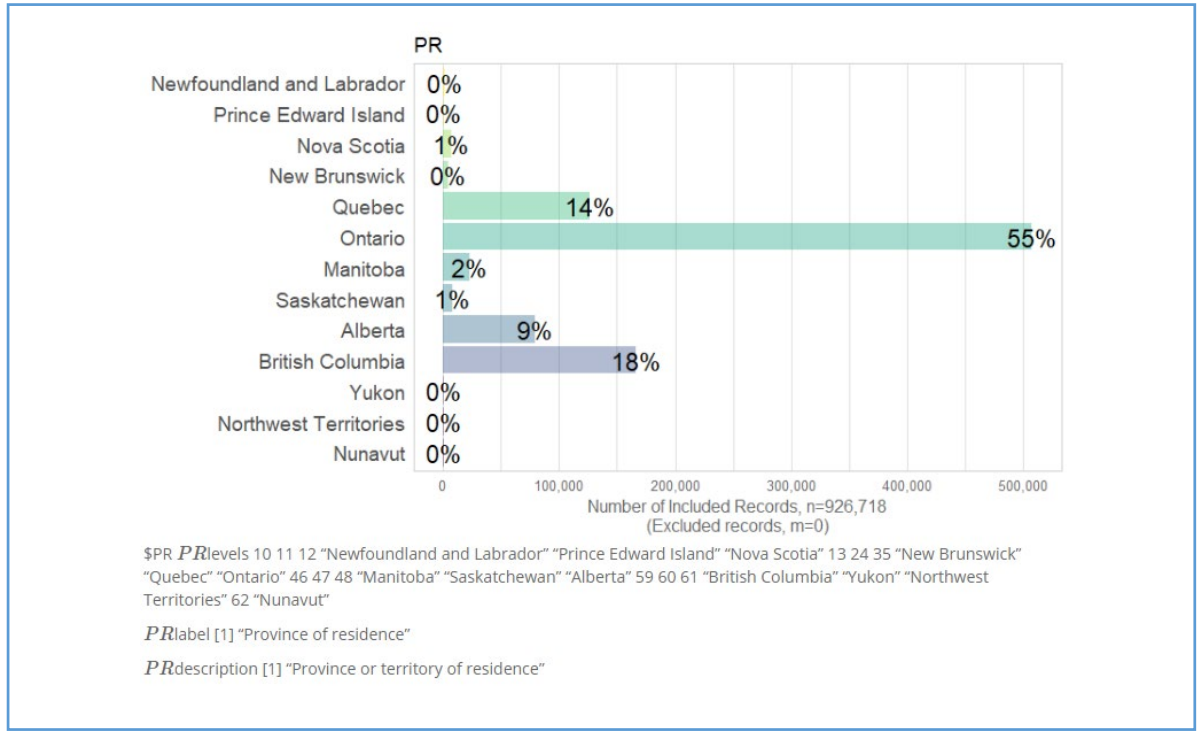
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

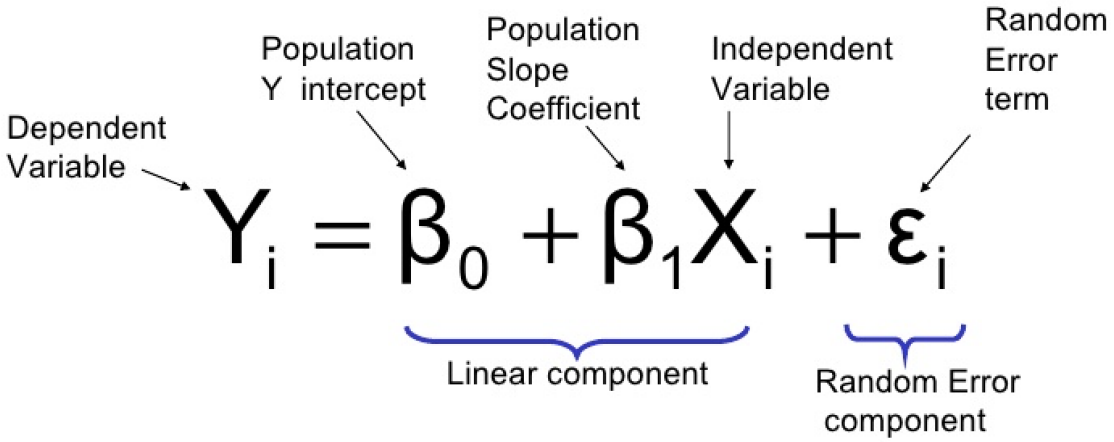
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Province of residence





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

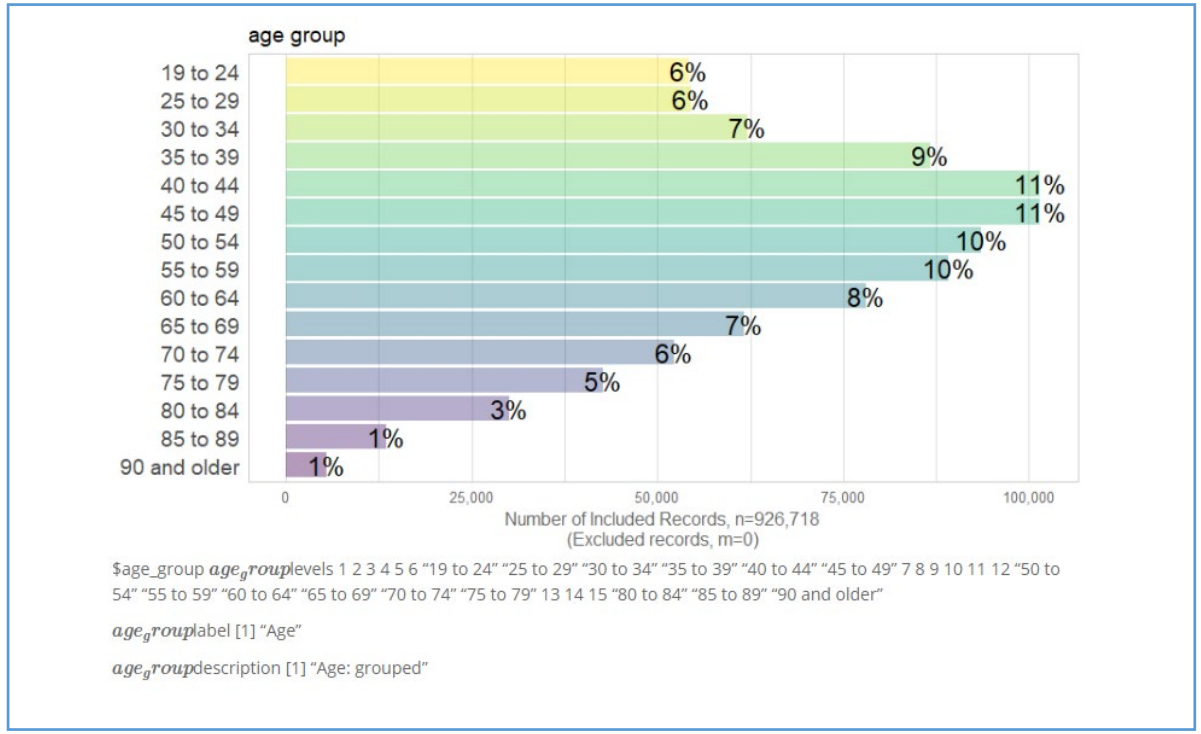
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

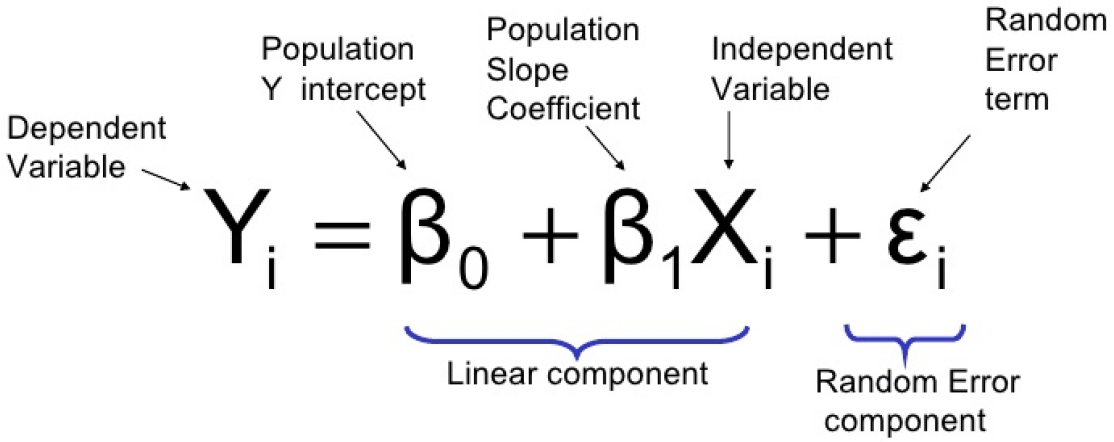
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

5-year age category





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

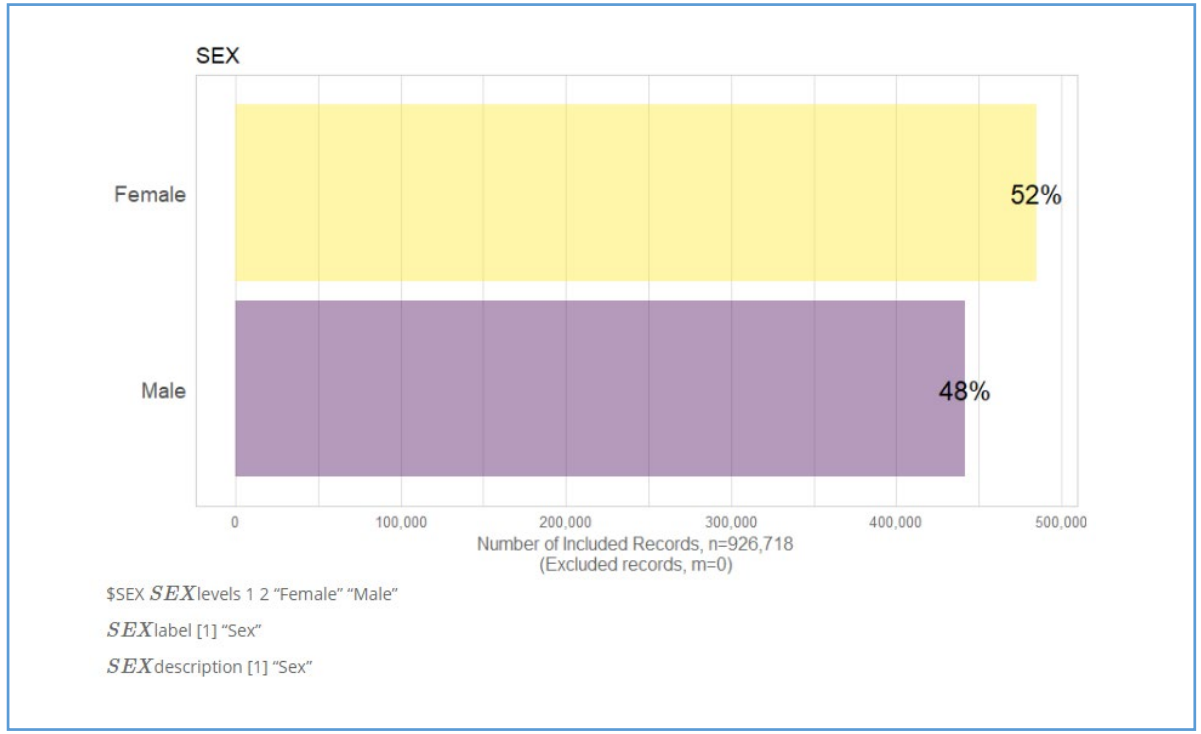
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

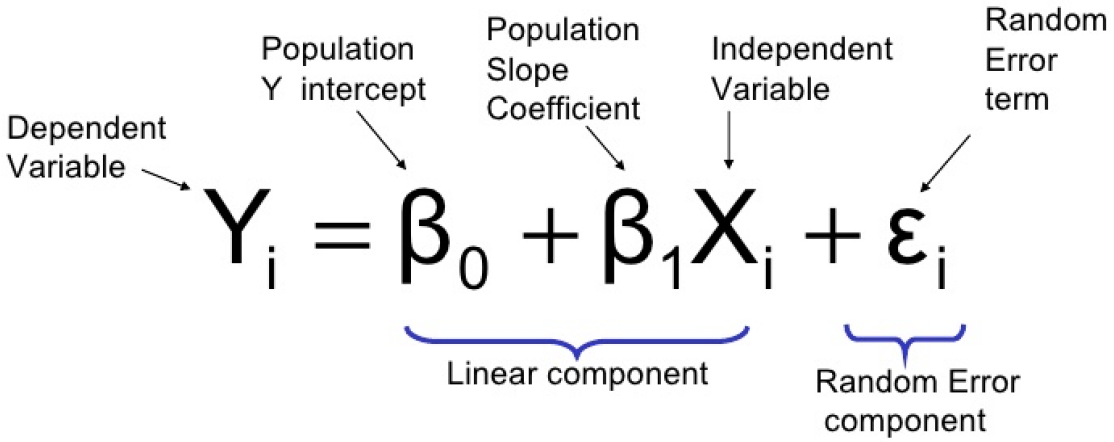
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Sex





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

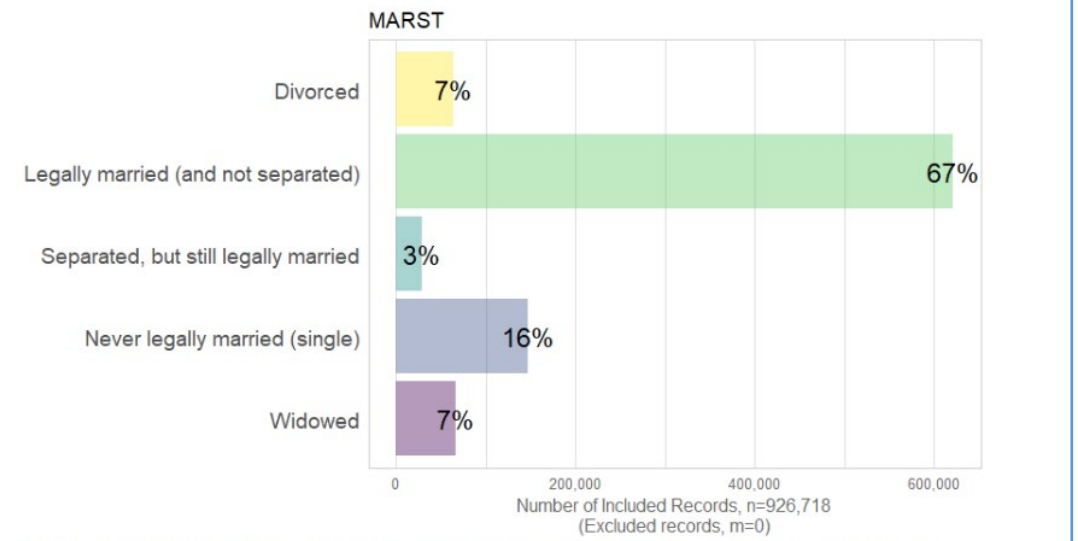
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Marital Status

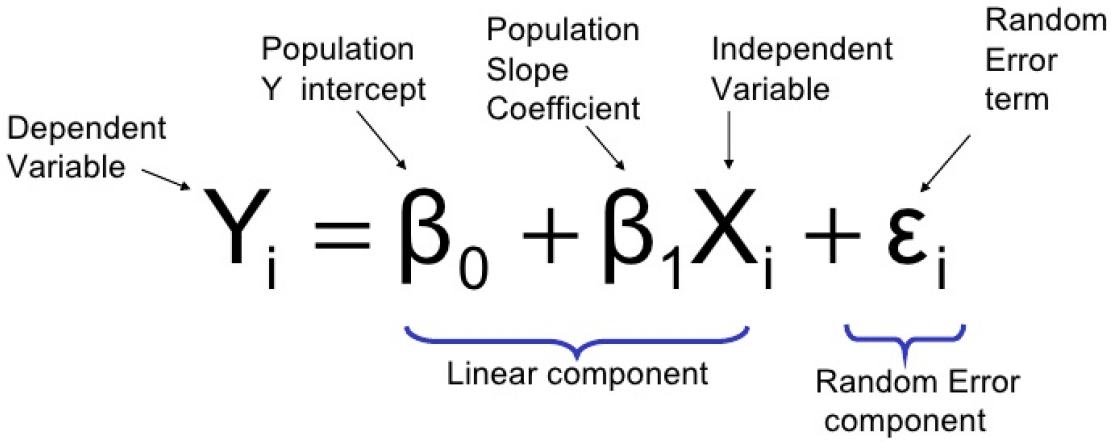
```
# because `still legaly married` is more legal than human
,marital = car::recode(
  MARST, "
  'Divorced'                              = 'sep_divorced'
  ;'Legally married (and not separated)'   = 'mar_cohab'
  ;'Separated, but still legally married'  = 'sep_divorced'
  ;'Never legally married (single)'        = 'single'
  ;'Widowed'                               = 'widowed'
  ")
,marital = factor(marital, levels = c(
  "sep_divorced","widowed","single","mar_cohab"))
```



$MARST $MARST$ levels 1 2 "Divorced" "Legally married (and not separated)" 3 4 "Separated, but still legally married" "Never legally married (single)" 5 "Widowed"

$MARST$ label [1] "Marital status"

$MARST$ description [1] "Marital Status: Refers to the legal marital status of the person."



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

0.1 Modeling form

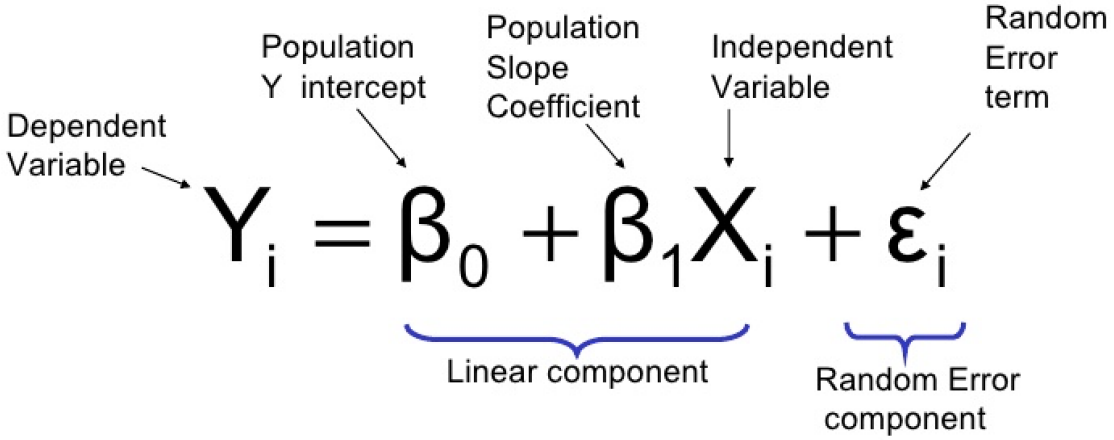
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Highest Degree

```
# because more than 5 categories is too fragmented
,educ5 = car::recode(
  HCDD, "
  'None'                                                                                    = 'less then high school'
  ;'High school graduation certificate or equivalency certificate'                          = 'high school'
  ;'Other trades certificate or diploma'                                                    = 'high school'
  ;'Registered apprenticeship certificate'                                                  = 'high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 3 months to less than 1 year' = 'college'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 1 year to 2 years'           = 'college'
  ;'College, CEGEP or other non-university certificate or diploma from a program of more than 2 years'           = 'college'
  ;'University certificate or diploma below bachelor level'                                = 'college'
  ;'Bachelors degree'                                                                       = 'college'
  ;'University certificate or diploma above bachelor level'                                = 'graduate'
  ;'Degree in medicine, dentistry, veterinary medicine or optometry'                        = 'graduate'
  ;'Masters degree'                                                                         = 'graduate'
  ;'Earned doctorate degree'                                                                = 'Dr.'
  ")
,educ5 = factor(educ5, levels = c(
  "less then high school"
  ,"high school"
  ,"college"
  ,"graduate"
  ,"Dr."
)
)
```

```
ds1 %>% group_by(educ5) %>% summarize(n = n())
```

```
# A tibble: 5 x 2
  educ5                    n
  <fct>                <int>
1 less then high school  902326
2 high school           1587347
3 college               1555485
4 graduate               269945
5 Dr.                     31546
```



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

0.1 Modeling form

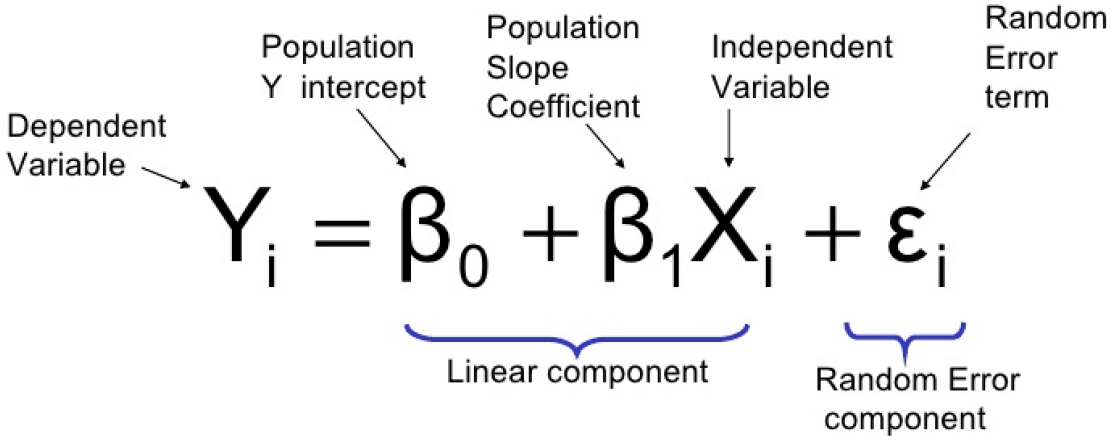
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Highest Degree

```
# because even only 5 may be too granular for our purposes
,educ3 = car::recode(
  HCDD, "
  'None'                                                                                                  = 'less than high school'
  ;'High school graduation certificate or equivalency certificate'                                        = 'high school'
  ;'Other trades certificate or diploma'                                                                  = 'high school'
  ;'Registered apprenticeship certificate'                                                                = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 3 months to less than 1 year' = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 1 year to 2 years'    = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of more than 2 years'    = 'more than high school'
  ;'University certificate or diploma below bachelor level'                                               = 'more than high school'
  ;'Bachelors degree'                                                                                      = 'more than high school'
  ;'University certificate or diploma above bachelor level'                                               = 'more than high school'
  ;'Degree in medicine, dentistry, veterinary medicine or optometry'                                       = 'more than high school'
  ;'Masters degree'                                                                                        = 'more than high school'
  ;'Earned doctorate degree'                                                                               = 'more than high school'
  ")
,educ3 = factor(educ3, levels = c(
  "less than high school"
  , "high school"
  , "more than high school"
)
)
```

```
# # because we want/need to inspect newly created variables
ds1 %>% group_by(educ3) %>% summarize(n = n())
```

```
# A tibble: 3 x 2
  educ3                    n
  <fct>                <int>
1 less than high school  902326
2 high school           1403807
3 more than high school 2040516
```



Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

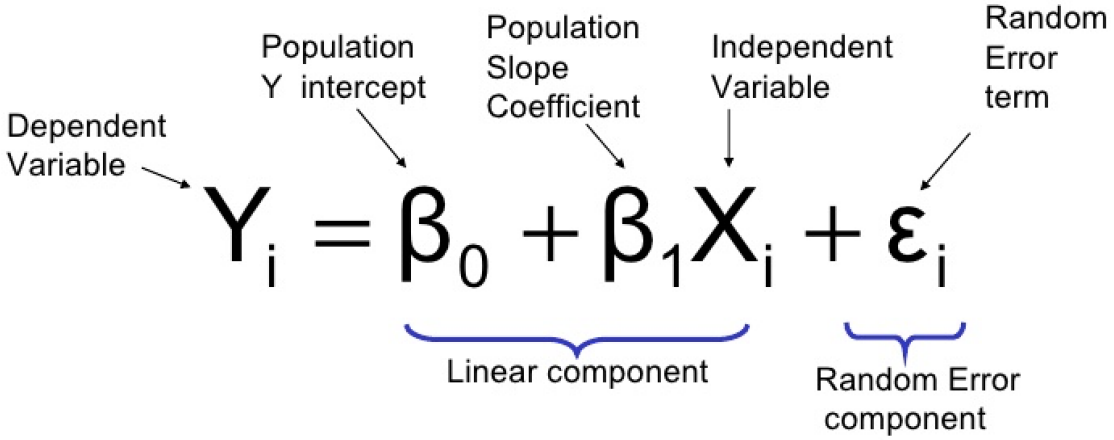
A. Graphing Technique

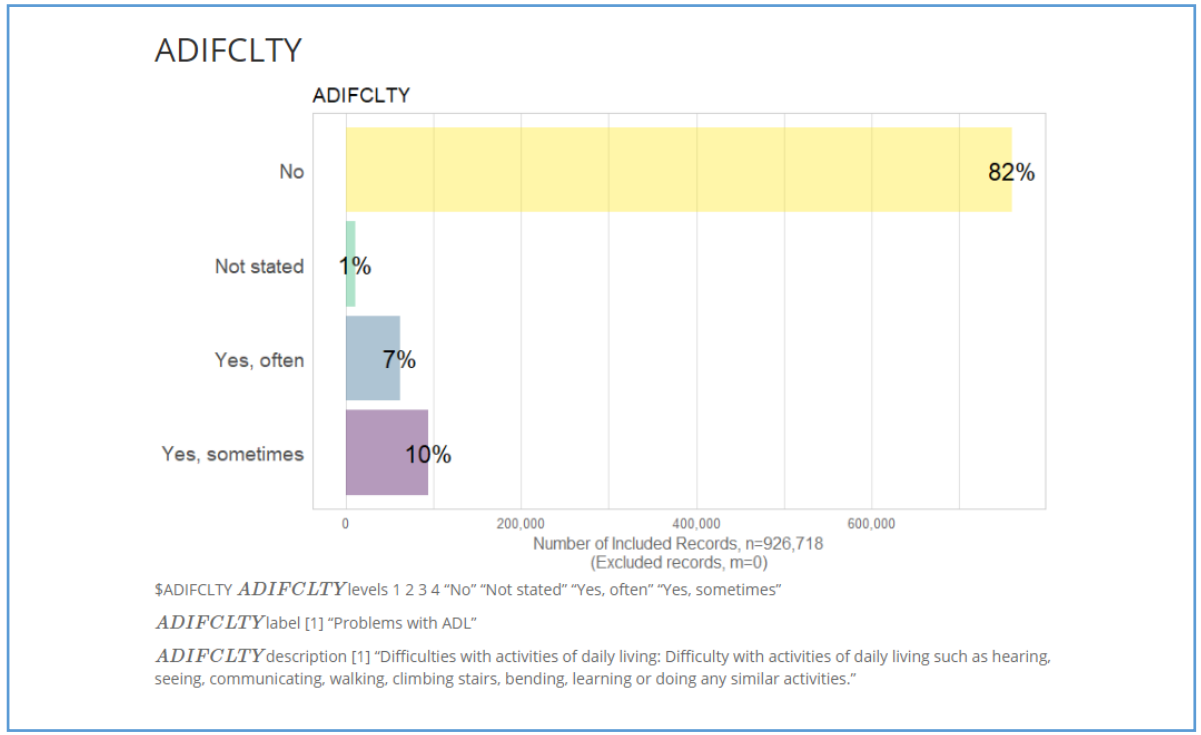
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Activities of Daily Living

```
# ADIFCLTY                "Problems with ADL" (physical & cognitive)
# DISABFL                 "Problems with ADL" (physical & social)
# because this is what counts practically
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                        DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
)
,poor_health = factor(poor_health, levels = c("TRUE","FALSE"))
```





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

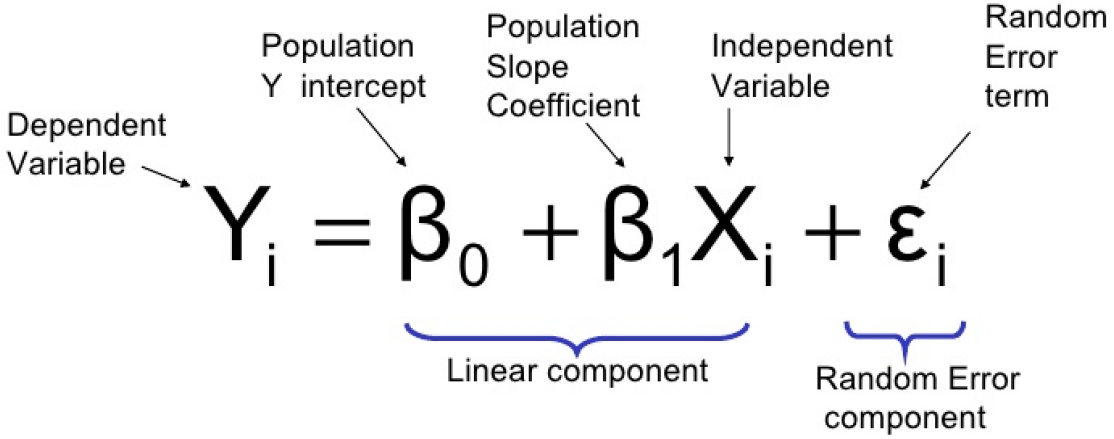
A. Graphing Technique

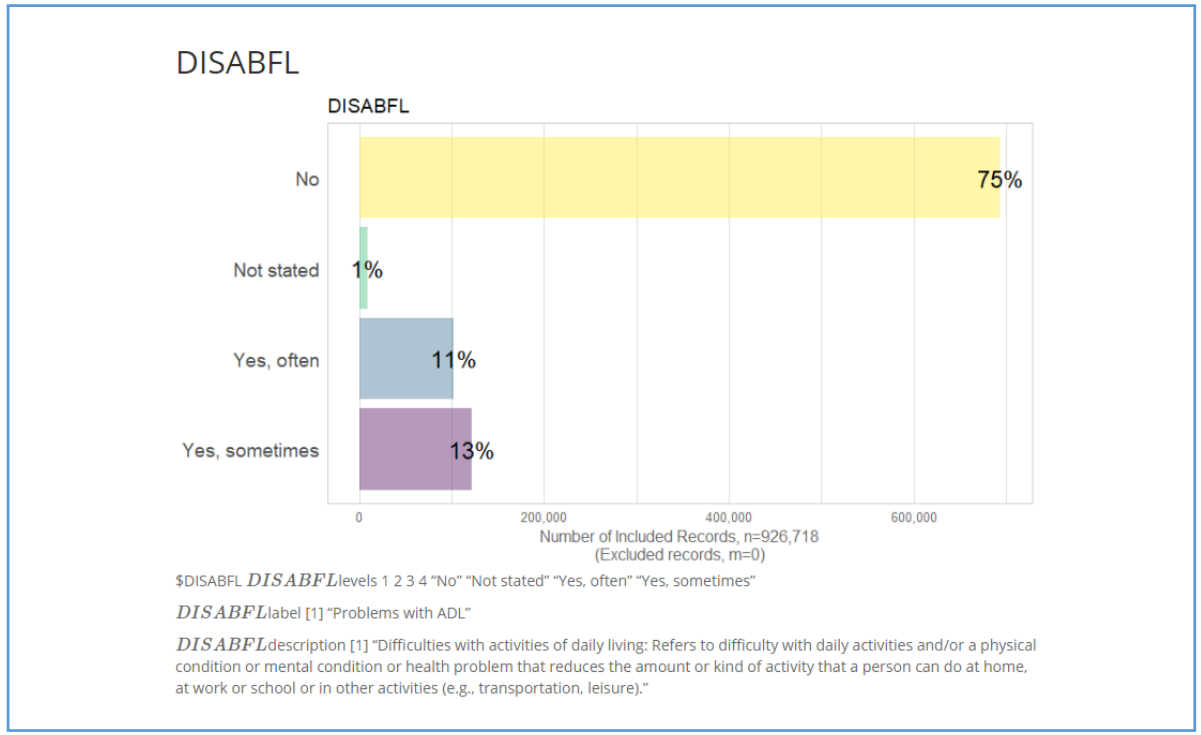
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Activities of Daily Living

```
# ADIFCLTY                "Problems with ADL" (physical & cognitive)
# DISABFL                 "Problems with ADL" (physical & social)
# because this is what counts practically
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                        DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
)
,poor_health = factor(poor_health, levels = c("TRUE","FALSE"))
```





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

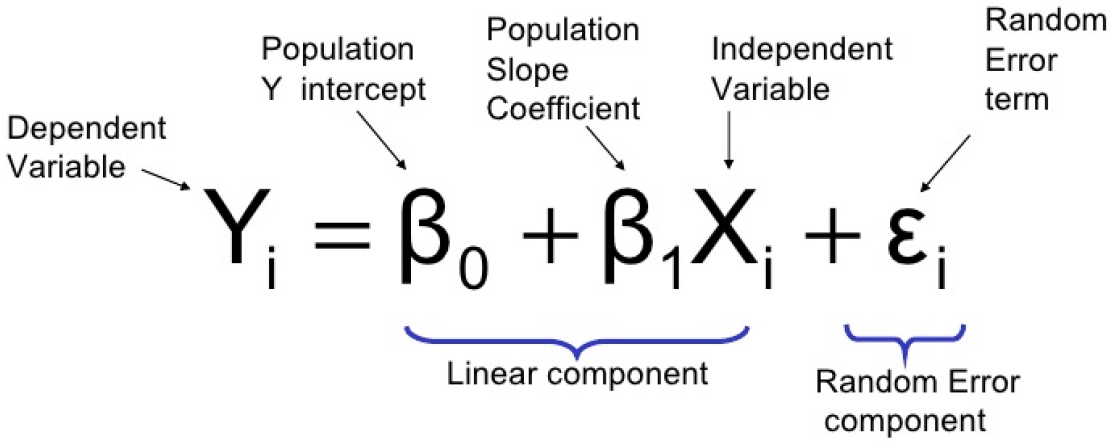
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

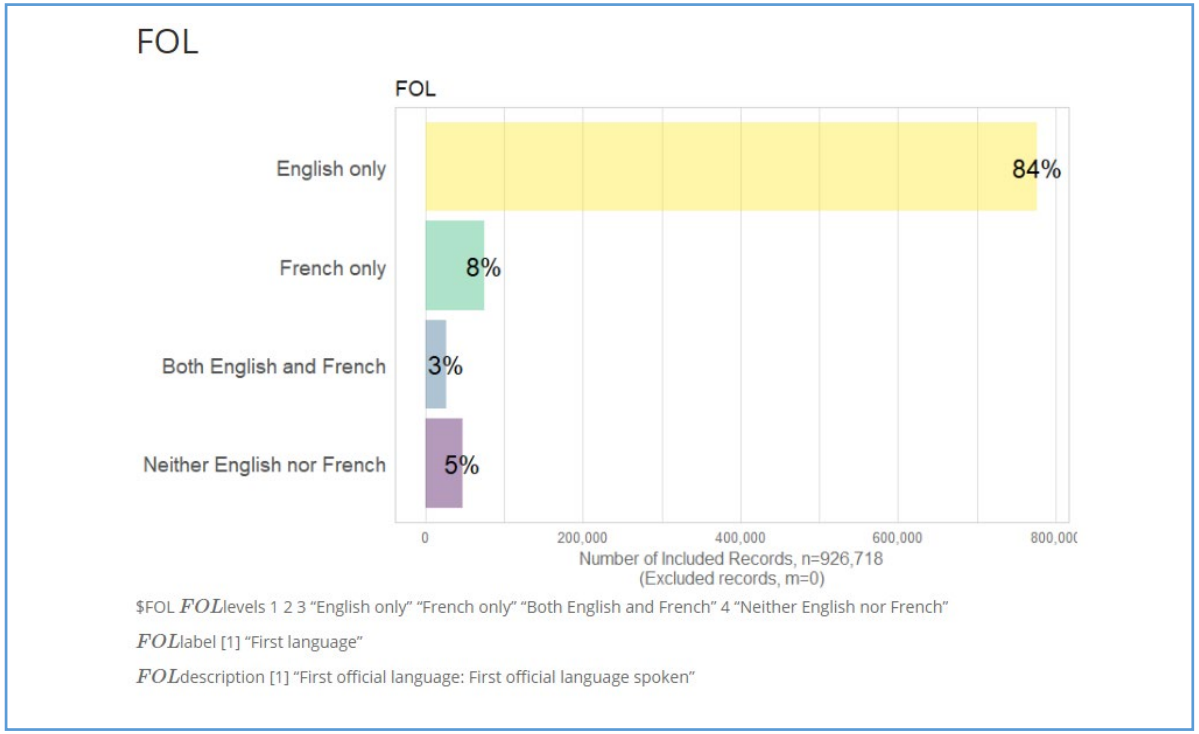
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

First Official Language



Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

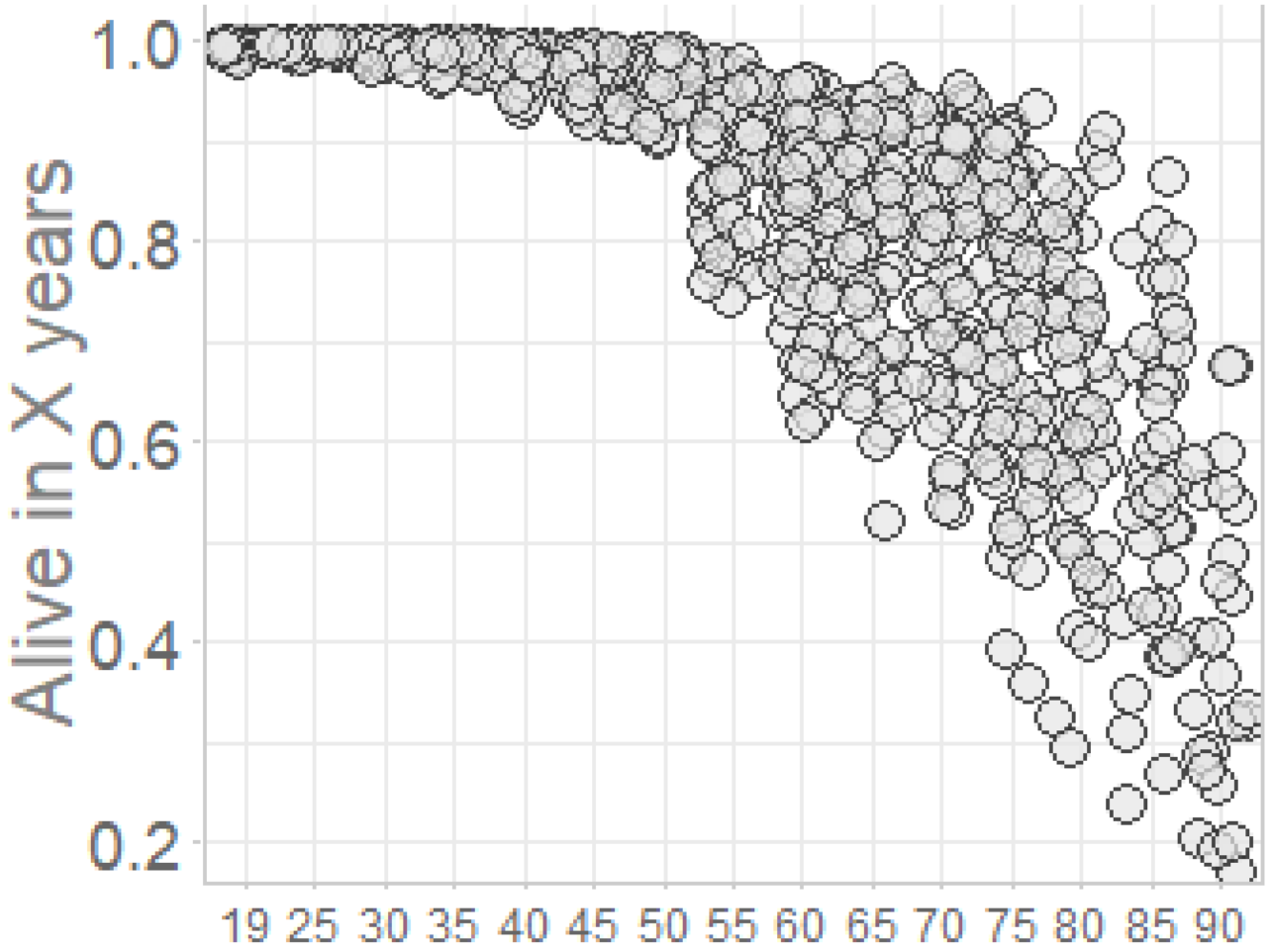
0.2 Graphical form

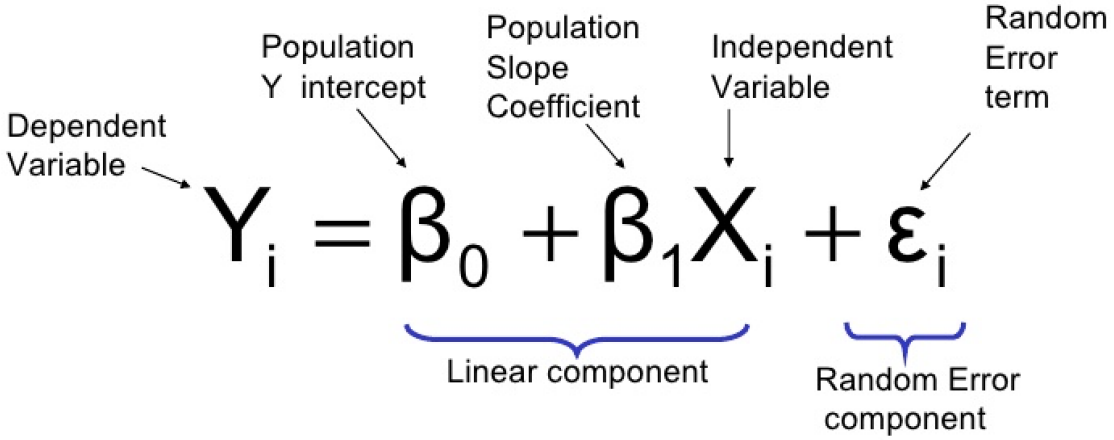
LEGEND

point = person
Y-axis = probability R is dead in X years
X-axis = age group (floor of 5-year category)

The higher the dot = the higher the chance to be alive in X years

Visualizing probability instead of log-odds because it is more intuitive





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

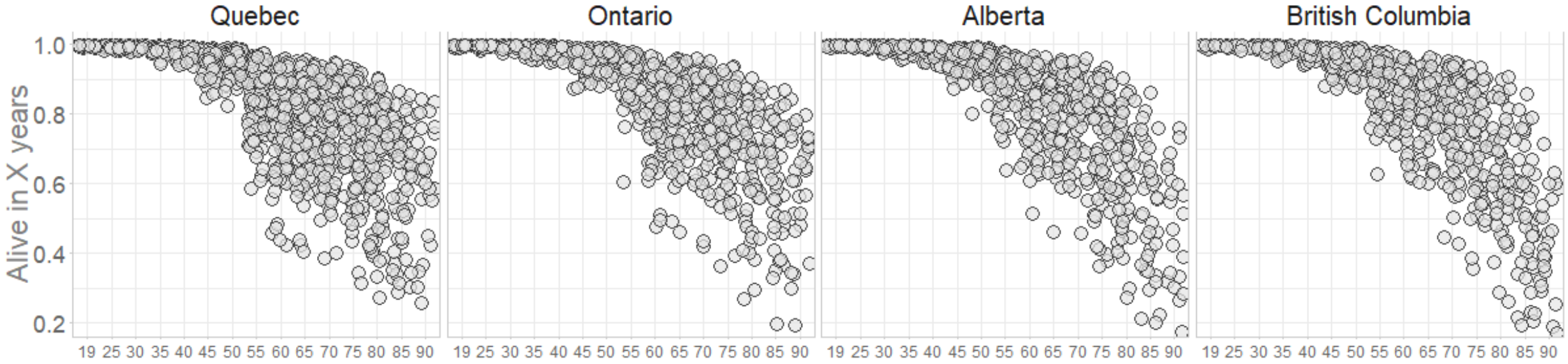
Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

0.2 Graphical form



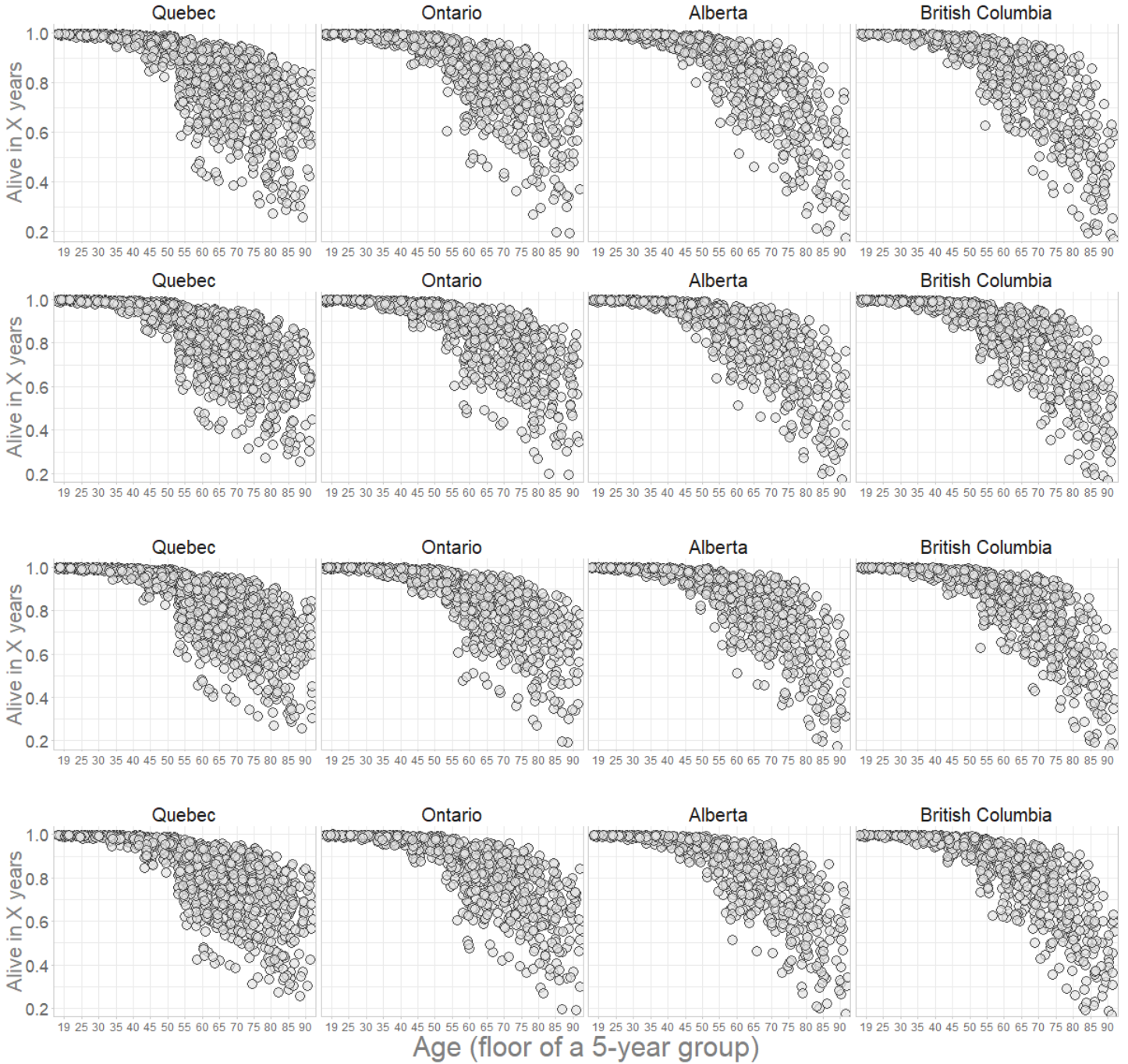
LEGEND

Facet = Province of residence

A. Graphing Technique

0.2 Graphical form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



LEGEND

Rows = duplicate of each other (for now).

Notice that FOL is not displayed

The book is ready for coloring

female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

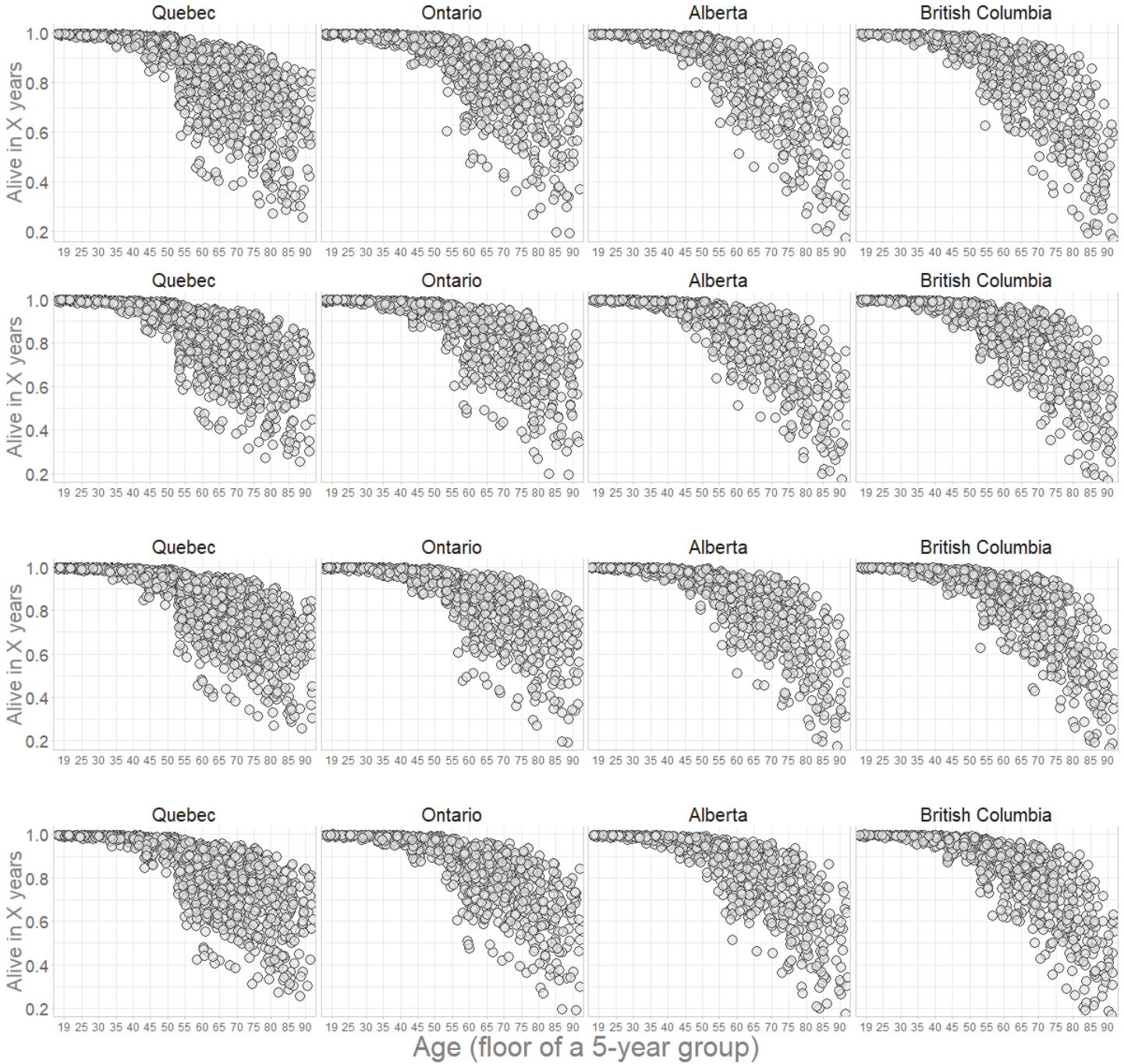
What should the "reference group" be for each predictor?

What do we expect based on existing research?

Informed expectation

Reference group

?



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

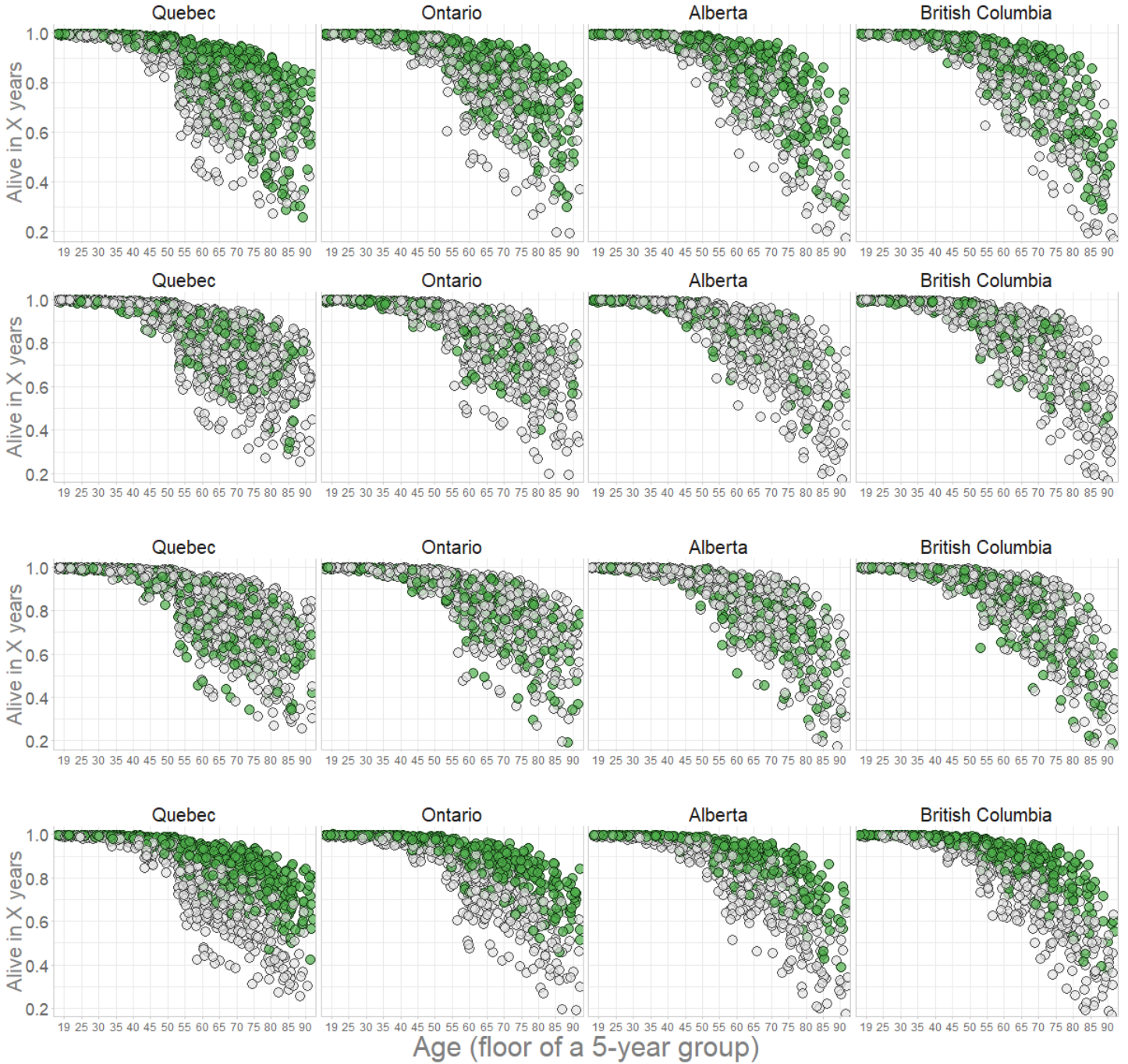
educ3
- less than high school
- high school
- more than high school

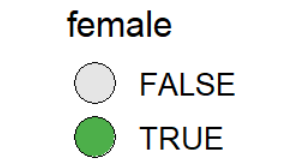
poor_health
- TRUE
- FALSE

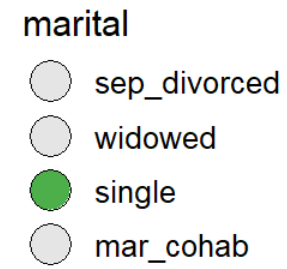
A. Graphing Technique

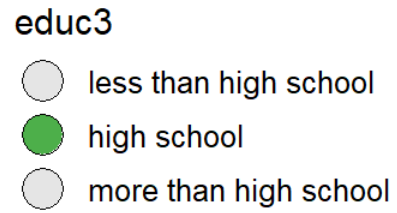
0.3 Coloring book

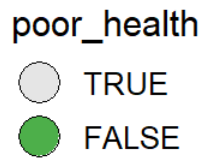
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL











Informed expectation

Reference group

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

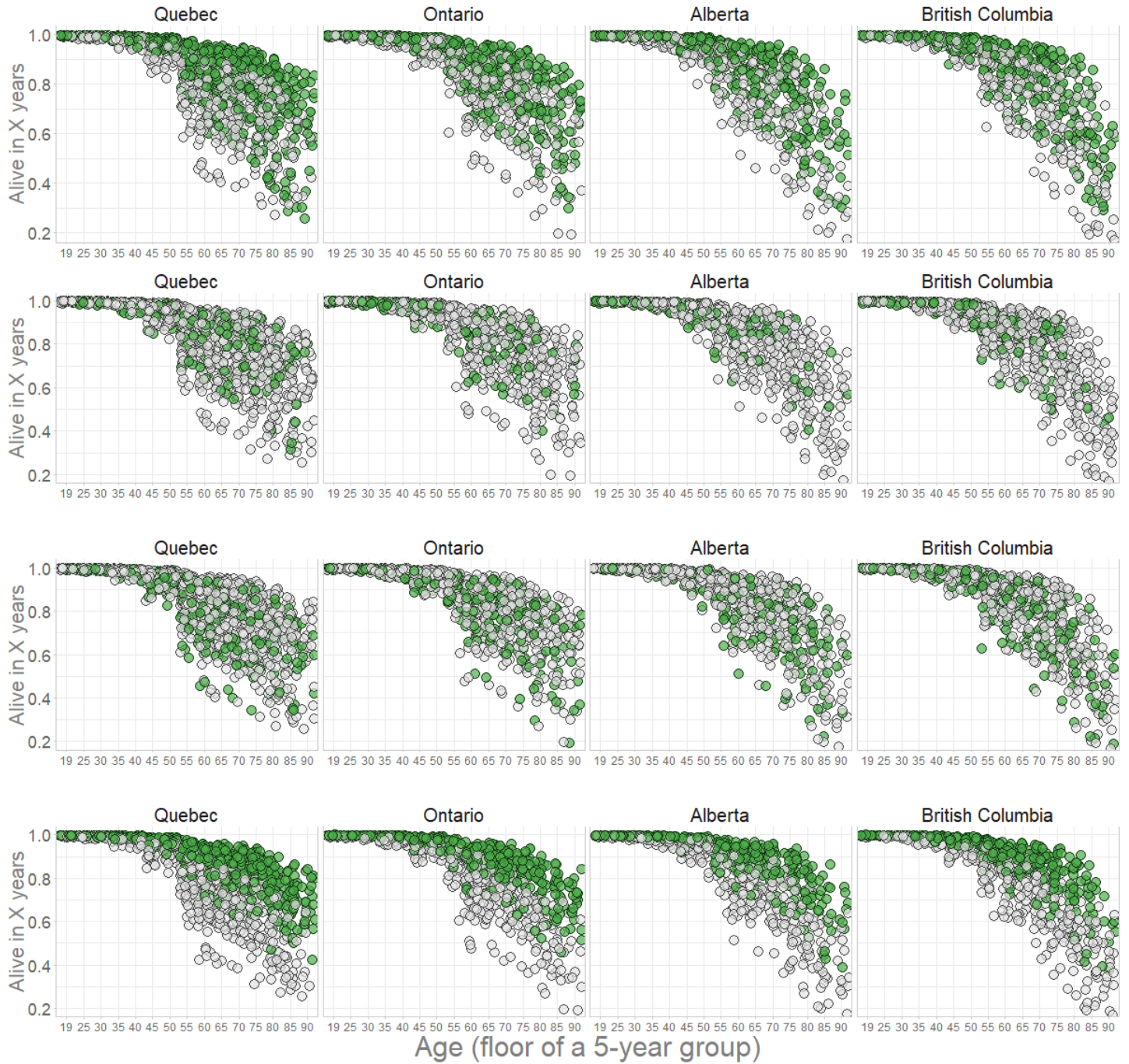
QUESTION

Compared to reference group, what levels of predictors are expected to **increase** the mortality risk?

Informed expectation

Moderately increased risk ?

Reference group



female
FALSE
TRUE

marital
sep_divorced
widowed
single
mar_cohab

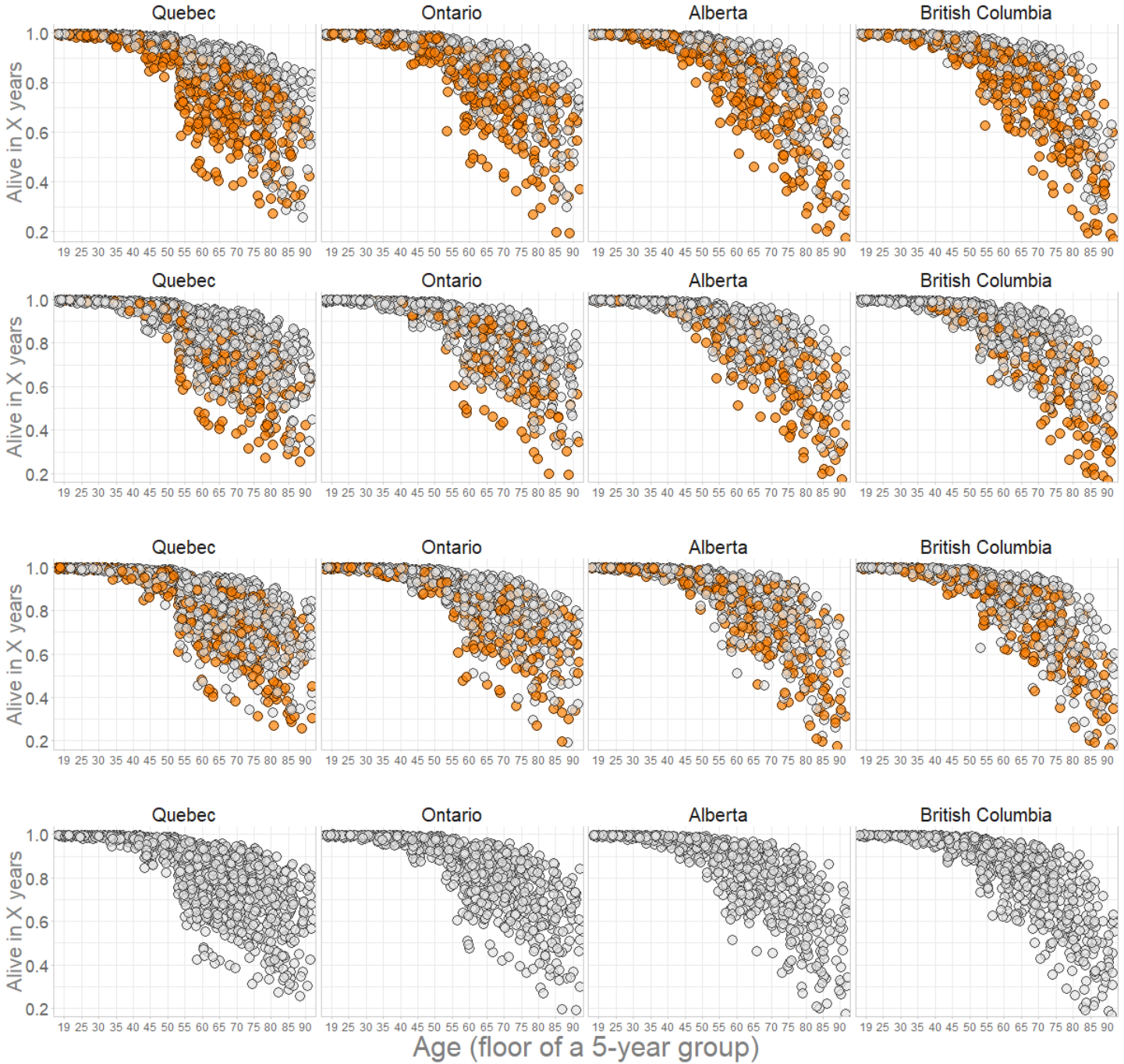
educ3
less than high school
high school
more than high school

poor_health
TRUE
FALSE

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

Informed expectation

Moderately increased risk

Reference group

A. Graphing Technique

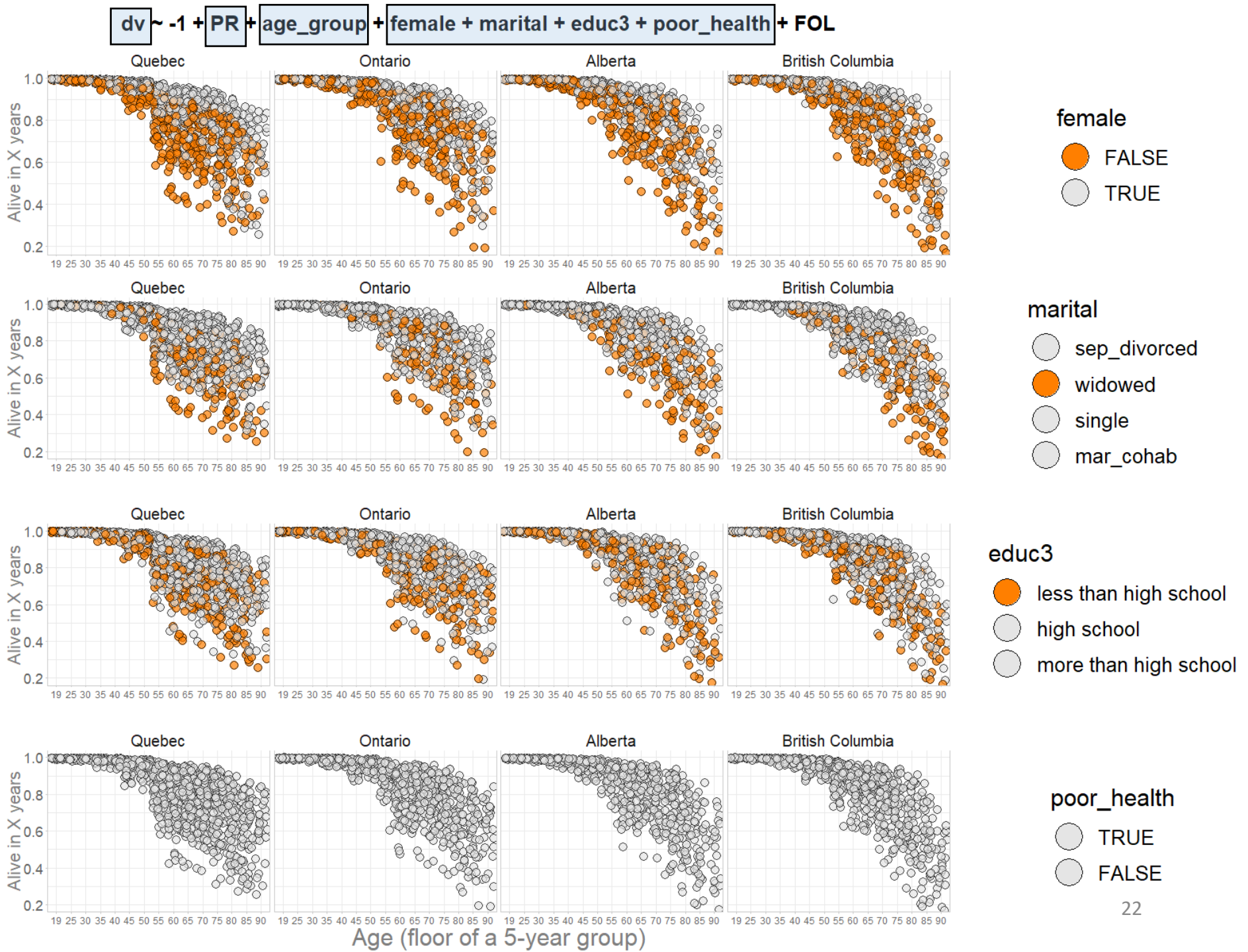
0.3 Coloring book

QUESTION

Compared to reference group, what levels of predictors are expected to **decrease** the mortality risk?

Informed expectation

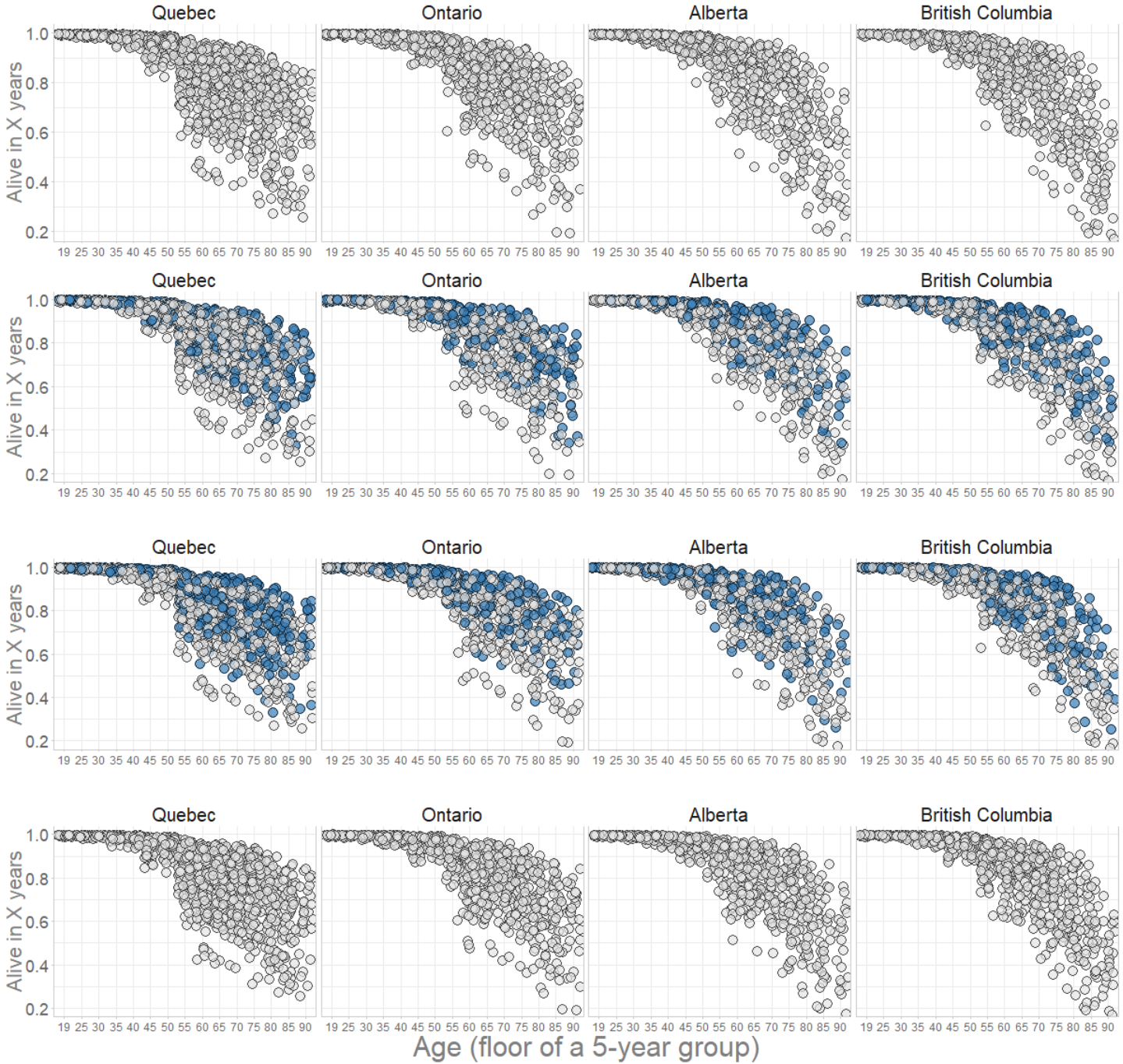
- Moderately increased risk
- Reference group
- Moderately decreased risk ?



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

Informed expectation

Moderately increased risk

Reference group

Moderately decreased risk

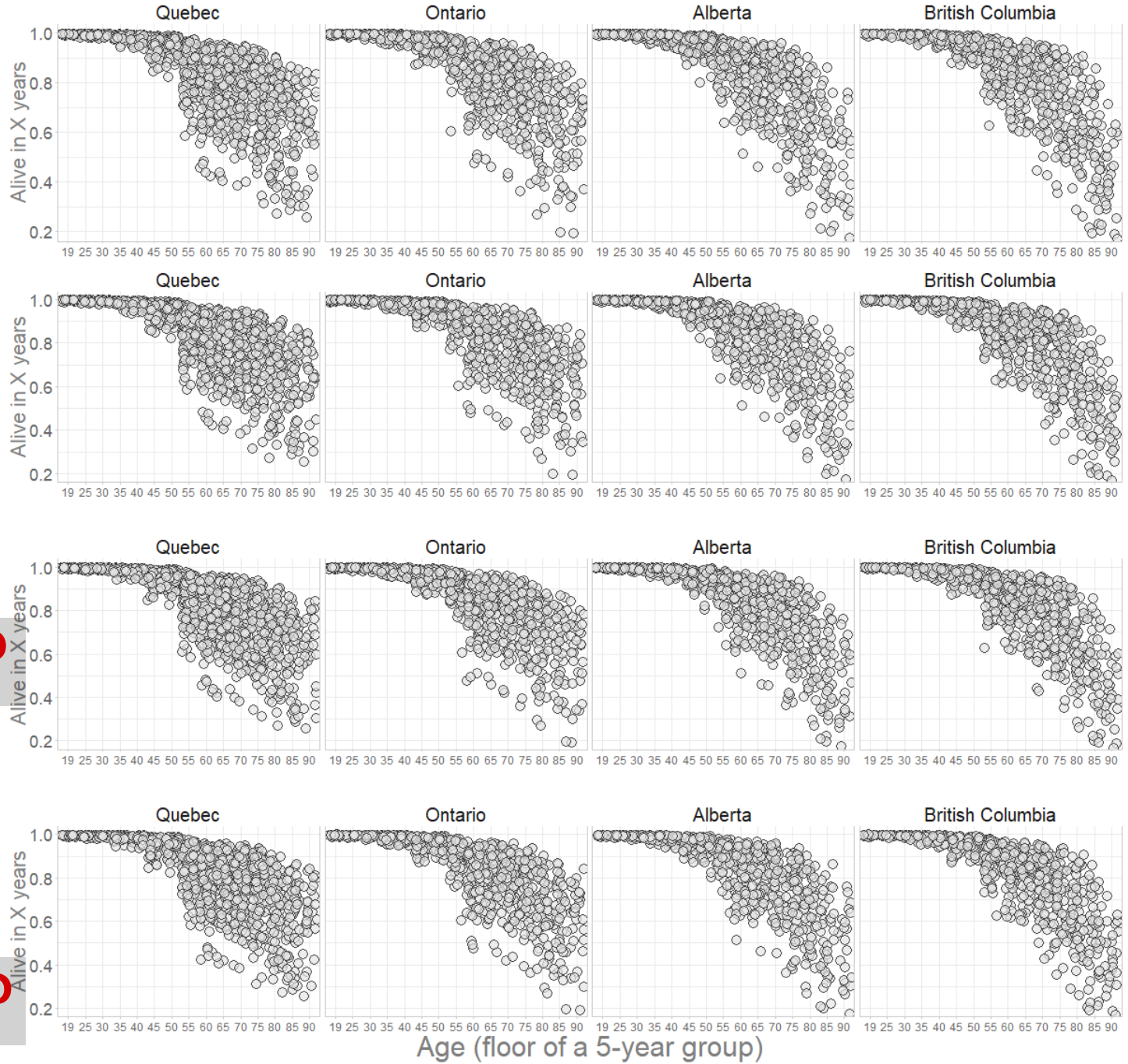
A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

What levels of predictors are expected to affect mortality risk drastically?



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

Informed expectation

- Substantially increased risk ?
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk ?

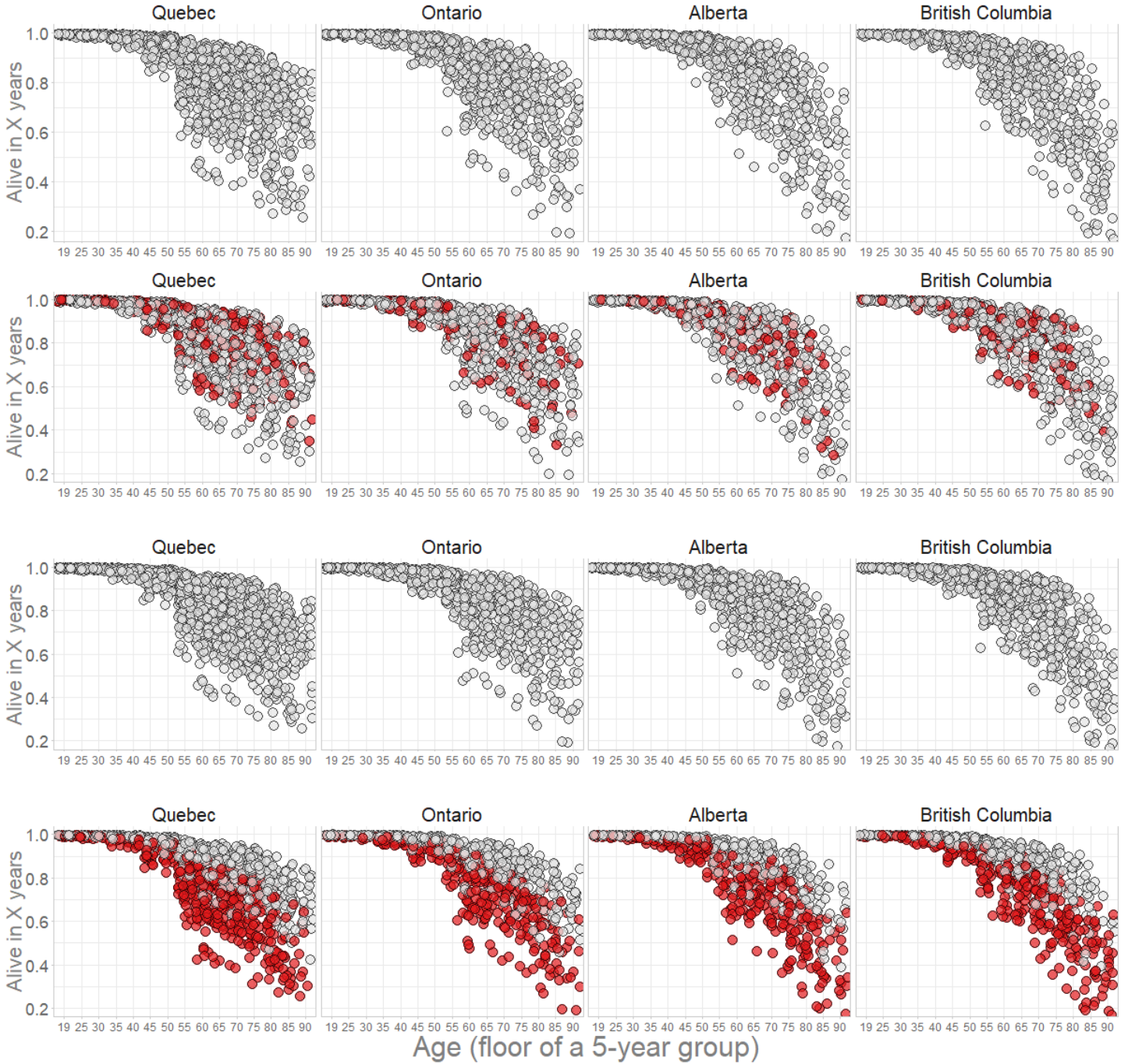
A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

What levels of predictors are expected to affect mortality risk drastically?



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk

A. Graphing Technique

0.3 Coloring book

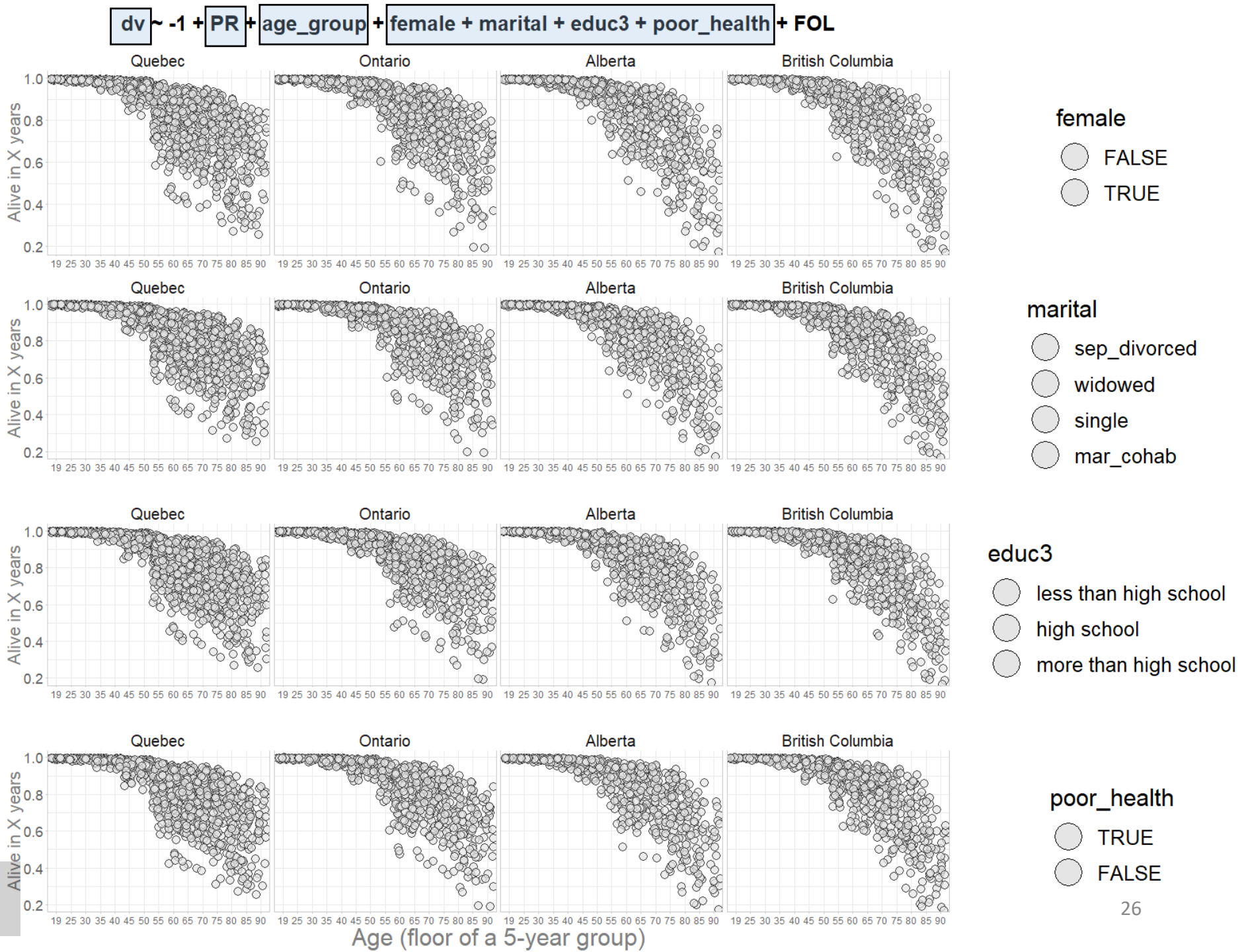
QUESTION

What levels of predictors are expected to affect mortality risk drastically?

No “very bad” and it’s ok.

Informed expectation

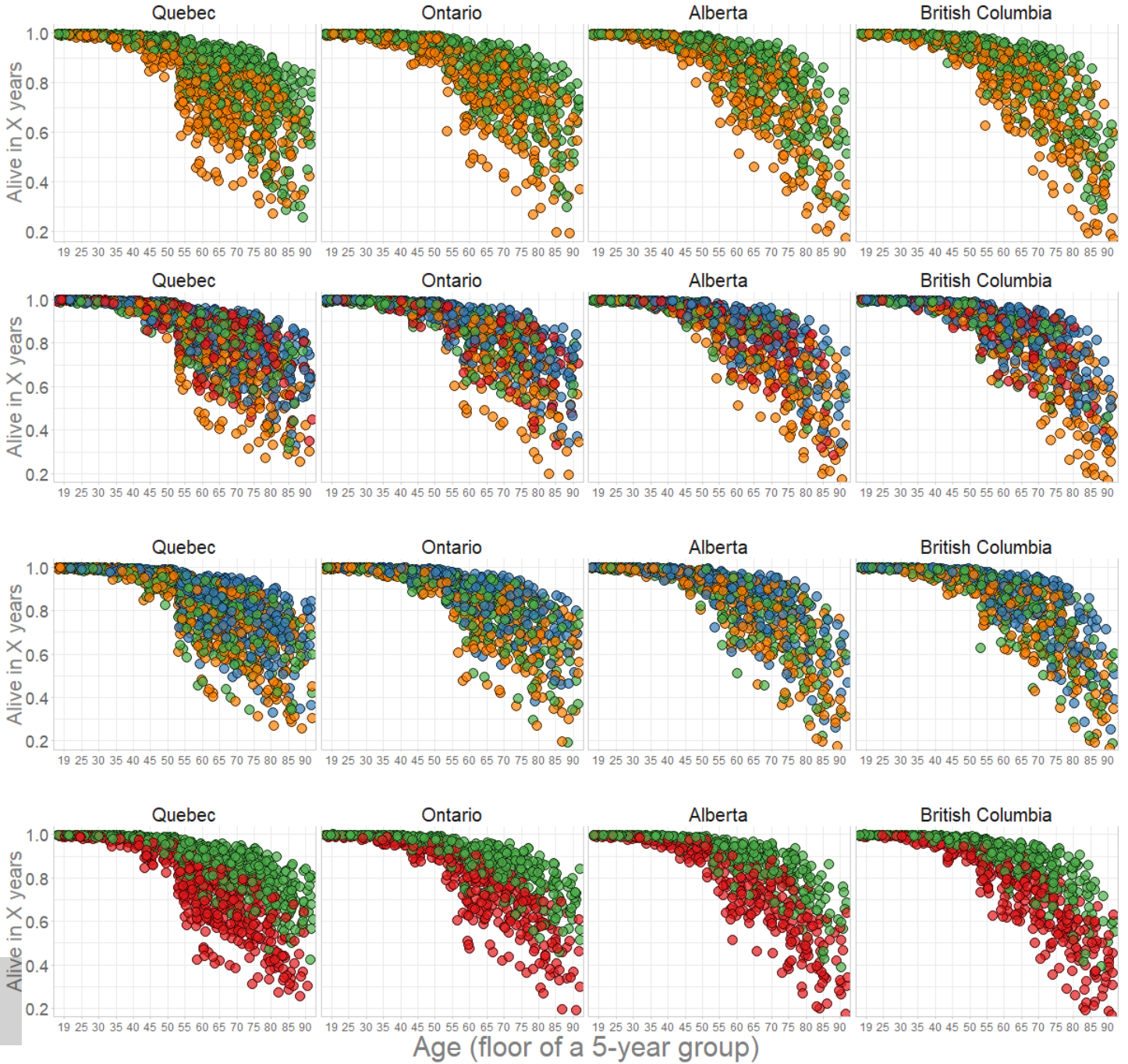
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

NOTICE

Plotting all colors at once may not be as informative as one would expect

May require too much tweaking to make useful

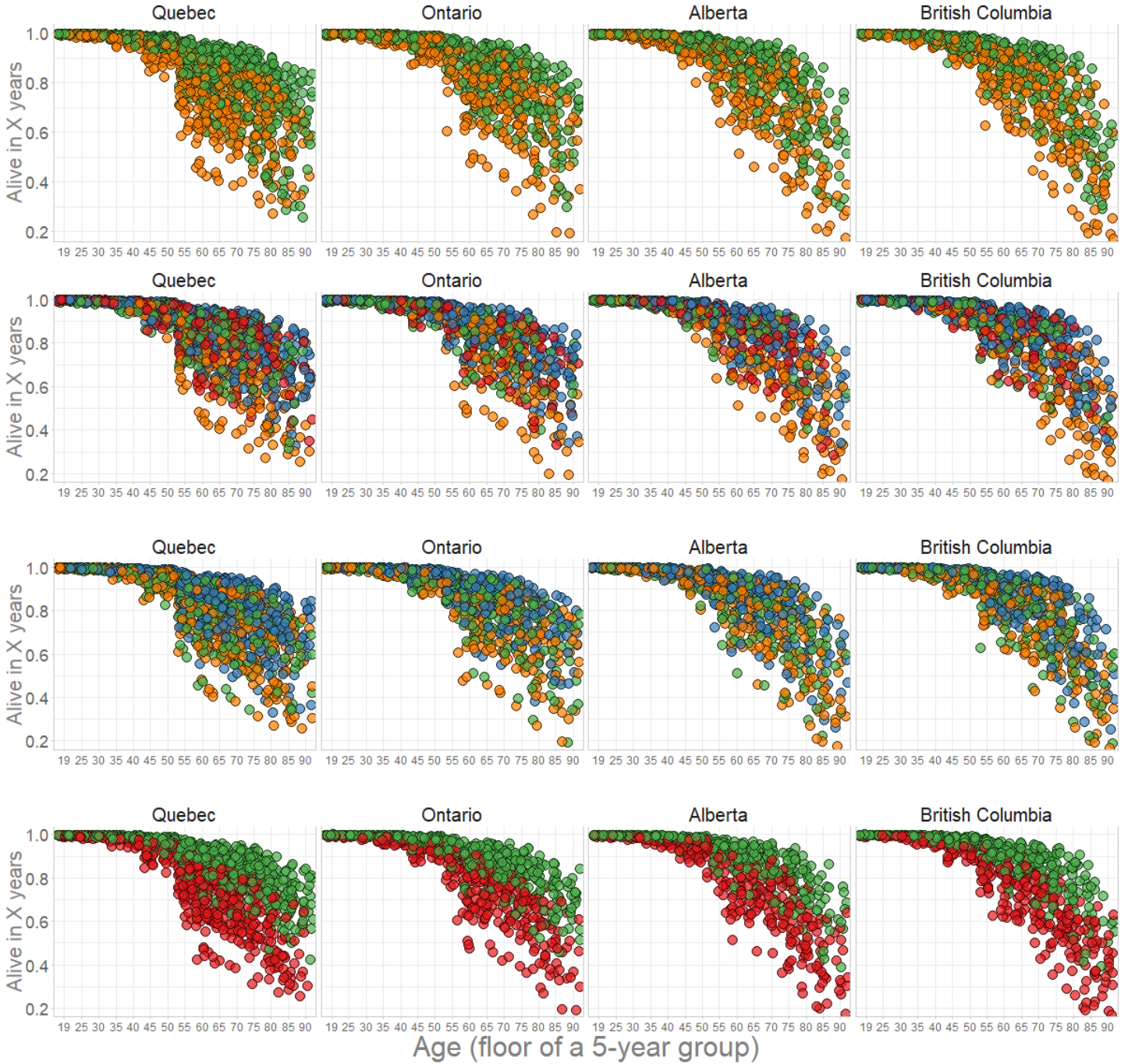
Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

NOTICE

Note all predictors are worth visualizing, some are there for control.

We can adjust what is being displayed

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk

A. Graphing Technique

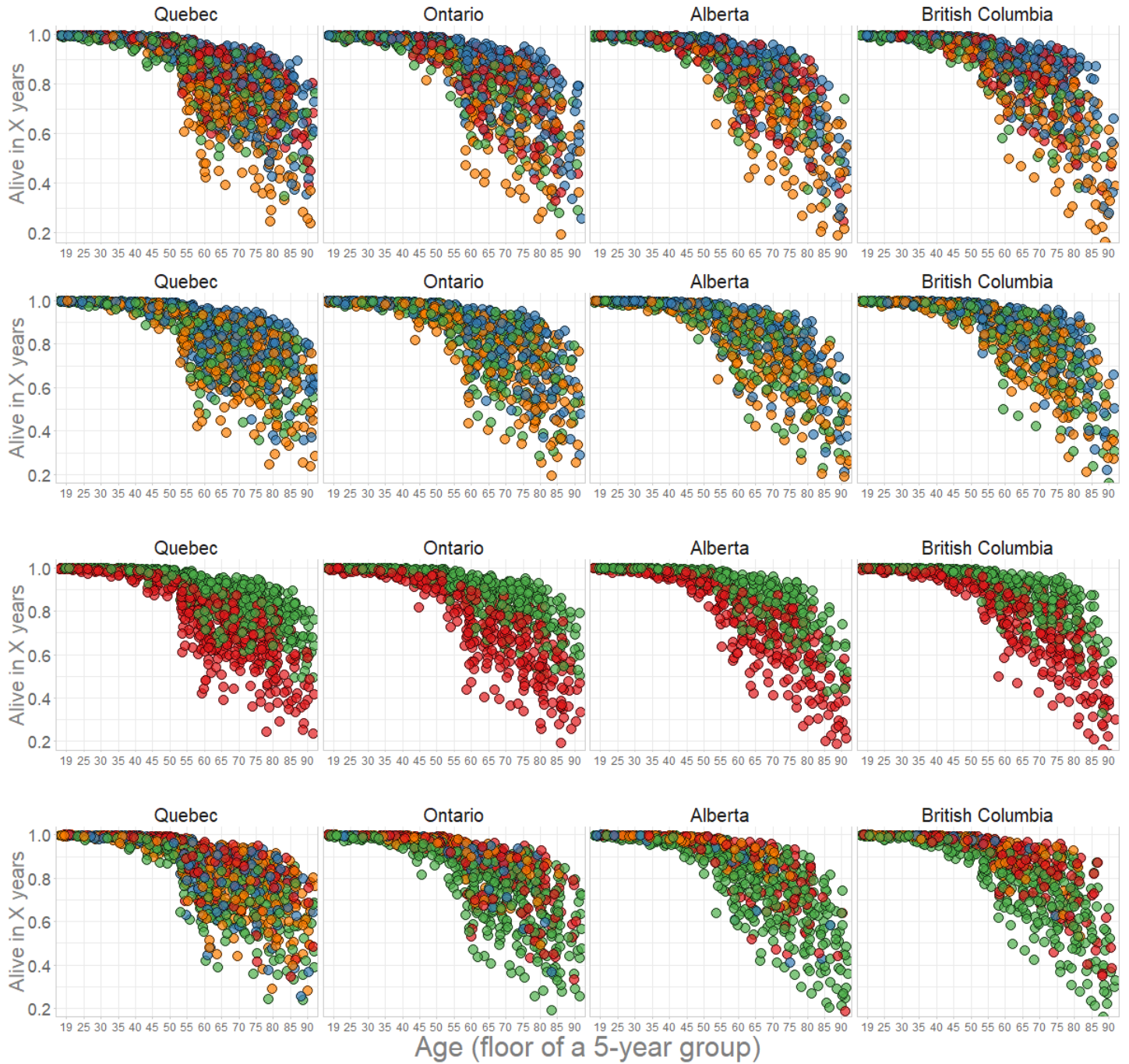
0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

NOTICE

Note all predictors are worth visualizing, some are there for control.

We can adjust what is being displayed



marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

FOL
- Neither English nor French
- French only
- English only
- Both English and French

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk

So how would you organize this production?

I cannot describe the workflow in the remaining time

But I can help you learn through reproduction

Here are some principles to keep in mind as you study the project

B. Workflow Highlights

1.0 “Let no one ignorant of geometry enter”: (my) scripts were written to be read by humans

How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

> ***Donald Knuth. "Literate Programming (1984)" in Literate Programming. CSLI, 1992, pg. 99.***
>
> I believe that the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be works of literature. Hence, my title: "Literate Programming."
>
> Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.

Source: http://www.literateprogramming.com/

Expect to read scripts

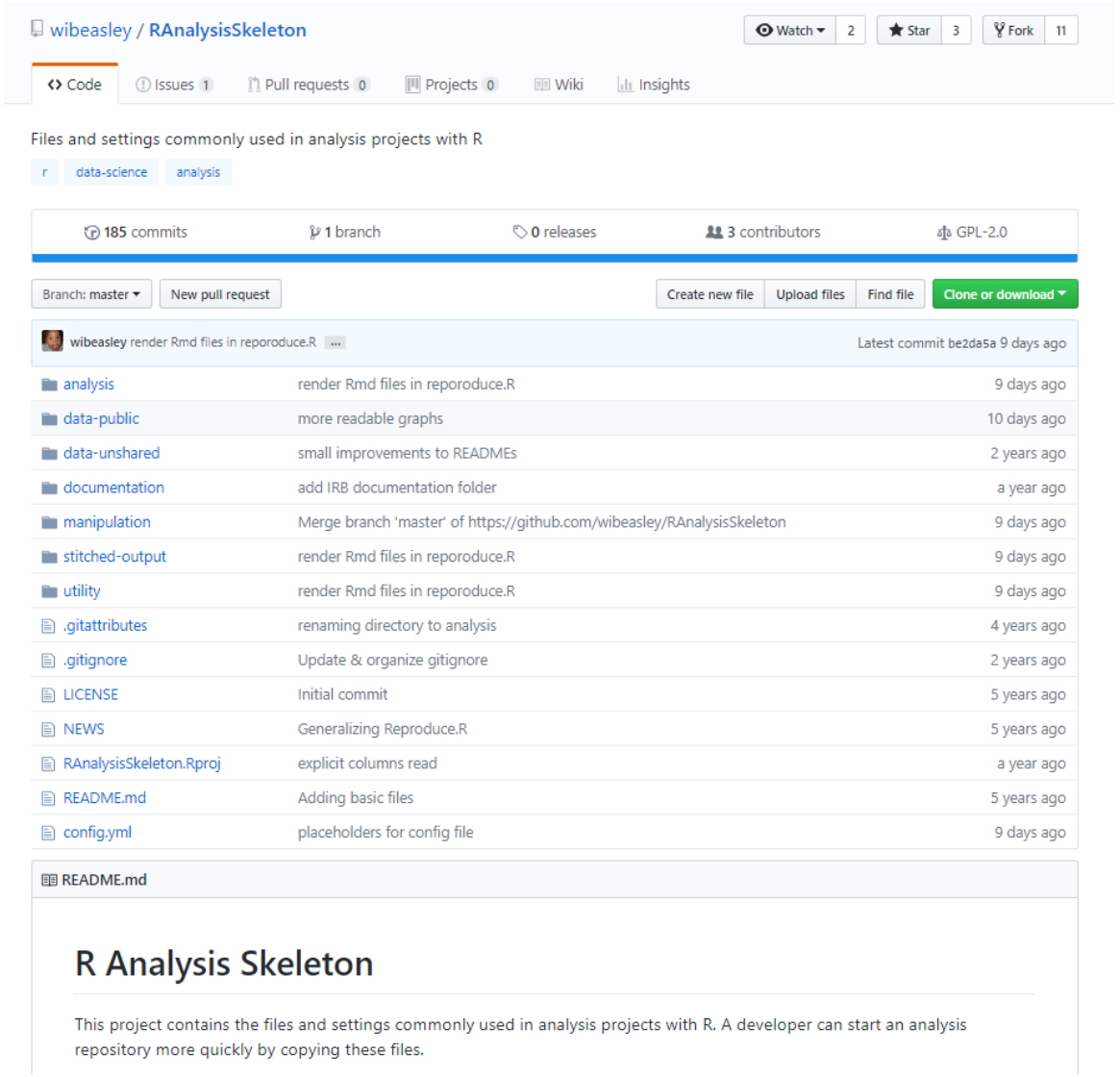
Main README should provide a map

https://github.com/andkov/ipdln-2018-hackathon/README.md

B. Workflow Highlights

1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects

Keep recognizable structure over projects



wibeasley / RAnalysisSkeleton

Watch 2 | Star 3 | Fork 11

Code | Issues 1 | Pull requests 0 | Projects 0 | Wiki | Insights

Files and settings commonly used in analysis projects with R

r | data-science | analysis

185 commits | 1 branch | 0 releases | 3 contributors | GPL-2.0

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

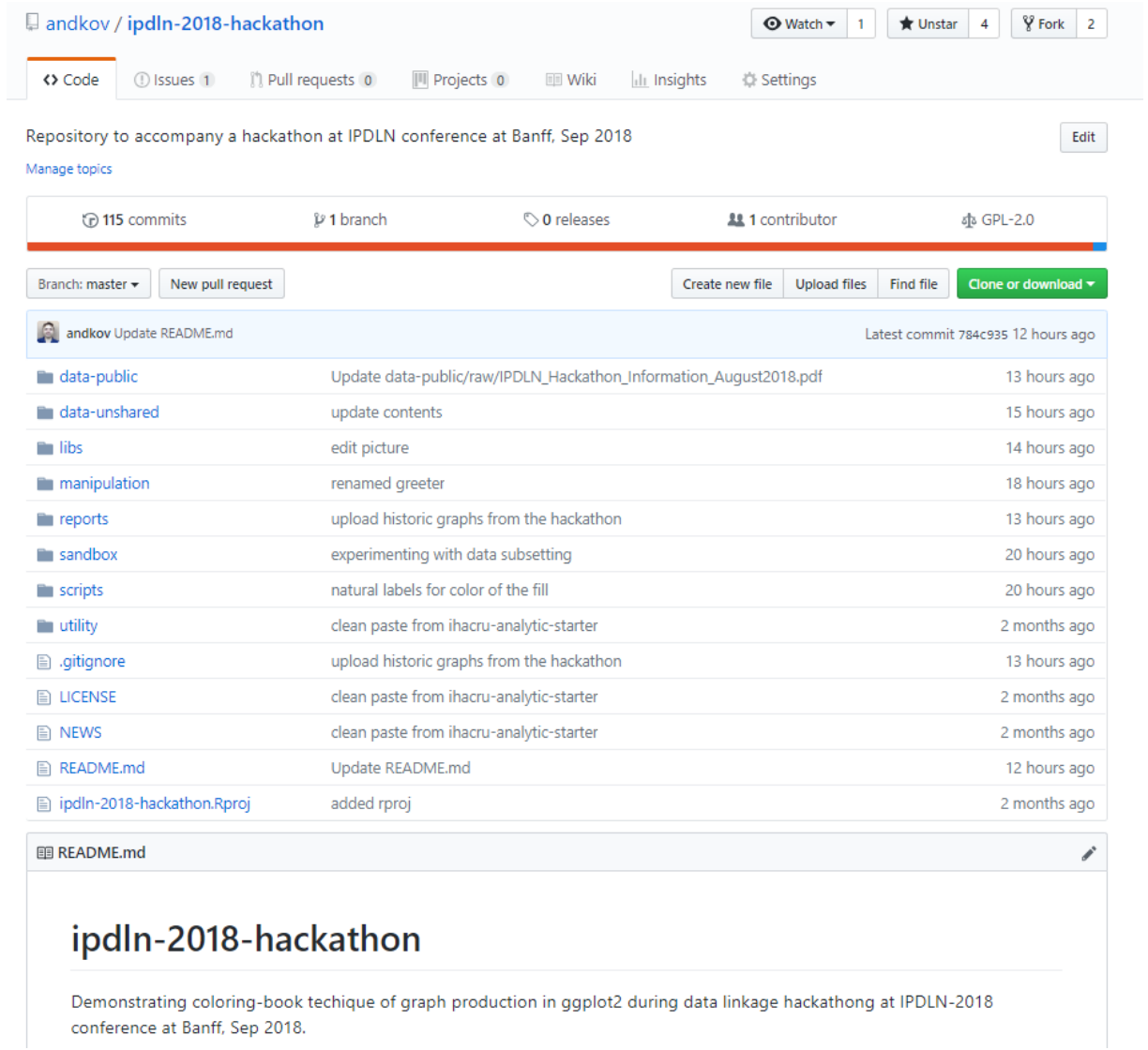
wibeasley render Rmd files in reporoduce.R ... Latest commit be2da5a 9 days ago

analysis	render Rmd files in reporoduce.R	9 days ago
data-public	more readable graphs	10 days ago
data-unshared	small improvements to READMEs	2 years ago
documentation	add IRB documentation folder	a year ago
manipulation	Merge branch 'master' of https://github.com/wibeasley/RAnalysisSkeleton	9 days ago
stitched-output	render Rmd files in reporoduce.R	9 days ago
utility	render Rmd files in reporoduce.R	9 days ago
.gitattributes	renaming directory to analysis	4 years ago
.gitignore	Update & organize gitignore	2 years ago
LICENSE	Initial commit	5 years ago
NEWS	Generalizing Reproduce.R	5 years ago
RAnalysisSkeleton.Rproj	explicit columns read	a year ago
README.md	Adding basic files	5 years ago
config.yml	placeholders for config file	9 days ago

README.md

R Analysis Skeleton

This project contains the files and settings commonly used in analysis projects with R. A developer can start an analysis repository more quickly by copying these files.



andkov / ipdln-2018-hackathon

Watch 1 | Unstar 4 | Fork 2

Code | Issues 1 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Repository to accompany a hackathon at IPDLN conference at Banff, Sep 2018

Edit

Manage topics

115 commits | 1 branch | 0 releases | 1 contributor | GPL-2.0

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

andkov Update README.md Latest commit 784c935 12 hours ago

data-public	Update data-public/raw/IPDLN_Hackathon_Information_August2018.pdf	13 hours ago
data-unshared	update contents	15 hours ago
libs	edit picture	14 hours ago
manipulation	renamed greeter	18 hours ago
reports	upload historic graphs from the hackathon	13 hours ago
sandbox	experimenting with data subsetting	20 hours ago
scripts	natural labels for color of the fill	20 hours ago
utility	clean paste from ihacru-analytic-starter	2 months ago
.gitignore	upload historic graphs from the hackathon	13 hours ago
LICENSE	clean paste from ihacru-analytic-starter	2 months ago
NEWS	clean paste from ihacru-analytic-starter	2 months ago
README.md	Update README.md	12 hours ago
ipdln-2018-hackathon.Rproj	added rproj	2 months ago

README.md

ipdln-2018-hackathon

Demonstrating coloring-book techique of graph production in ggplot2 during data linkage hackathong at IPDLN-2018 conference at Banff, Sep 2018.

Notice structural similarities

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

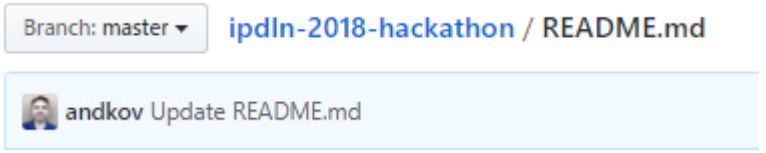
Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.



Try to keep tasks separate:
- Data cleaning
- Statistical modeling
- Graph production

Tasks are narratives to be told

Here are some examples

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of linked dynamic document

```
# declare where you will store the product of this script
path_save <- "./data-unshared/derived/ls_guide.rds"
```

```
POBDER <- list(
  "levels" = c(
    "1" = " Born in province of residence"
   ,"2" = " Born in another province"
   ,"3" = " Born outside Canada "
  )
  ,"label" = "Place of birth"
  ,"description"= "Place of birth: Indicates whether the respondent was born in the same province that they li
)
PR <- list(
  "levels" = c(
    "10" = "Newfoundland and Labrador"
   ,"11" = "Prince Edward Island"
   ,"12" = "Nova Scotia"
   ,"13" = "New Brunswick"
   ,"24" = "Quebec"
   ,"35" = "Ontario"
   ,"46" = "Manitoba"
   ,"47" = "Saskatchewan"
   ,"48" = "Alberta"
   ,"59" = "British Columbia"
   ,"60" = "Yukon"
   ,"61" = "Northwest Territories"
   ,"62" = "Nunavut"
  )
  ,"label" = "Province of residence"
  ,"description"= "Province or territory of residence"
)
```

```
# create vector with names
block_names <- c("demographic", "identity", "economic", "immigration","health")
item_names  <- c(demographic,   identity,   economic,   immigration,  health)
# create a list object to hold all available metadata
ls_guide            <- list()
ls_guide[["block"]] <- mget(block_names, envir = globalenv())
ls_guide[["item"]]  <- mget(item_names,  envir = globalenv())
```

```
# show components of this list object
ls_guide %>% lapply(names)
```

```
## $block
## [1] "demographic" "identity"    "economic"    "immigration" "health"
##
## $item
##  [1] "SEX"                  "age_group"
##  [3] "MARST"                "EFCNT_PP_R"
##  [5] "KID_group"            "PR"
##  [7] "FOL"                  "OLN"
##  [9] "DVISMIN"              "ABDERR"
## [11] "ABIDENT"              "HCDD"
## [13] "COWD"                 "NOCSBRD"
## [15] "TRMODE"               "LOINCA"
## [17] "LOINCB"               "d_licoratio_da_bef"
## [19] "RUINDFG"              "RPAIR"
## [21] "POBDER"               "DPOB11N"
## [23] "IMMDER"               "AGE_IMM_REVISED_group"
## [25] "YRIM_group"           "CITSM"
## [27] "GENSTPOB"             "ADIFCLTY"
## [29] "DISABFL"              "DISABIL"
## [31] "S_DEAD"               "COD1"
## [33] "COD1_CODES"           "COD2"
## [35] "COD2_CODES"
```

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of linked dynamic document

```
# link to the source of the location mapping
path_input_micro <- "./data-unshared/raw/ipdln_synth_final.csv"
path_input_meta  <- "./data-unshared/derived/ls_guide.rds"

# test whether the file exists / the link is good
testit::assert("File does not exist", base::file.exists(path_input_micro))
testit::assert("File does not exist", base::file.exists(path_input_meta))

# declare where you will store the product of this script
path_save <- "./data-unshared/derived/0-greeted.rds"
```

```
ds0      <- readr::read_csv(path_input_micro) %>% as.data.frame()
```

```
# basic inspection
ds0 %>% dplyr::glimpse(50)
```

```
## Observations: 4,346,649
## Variables: 34
## $ ABDERR_synth
## $ ABIDENT_synth
## $ ADIFCLTY_synth
## $ CITSM_synth
## $ COWD_synth
## $ DISABFL_synth
## $ DISABIL_synth
## $ DVISMIN_synth
## $ FOL_synth
## $ FPTIM_synth
## $ GENSTPOB_synth
## $ HCDD_synth
## $ IMMDER_synth
## $ LOINCA_synth
## $ LOINCB_synth
## $ MARST_synth
## $ NOCSBRD_synth
## $ OLN_synth
## $ POBDER_synth
## $ SEX_synth
## $ TRMODE_synth
## $ RPAIR_synth
## $ PR_synth
```

```
cat("Save results to ",path_save)
```

```
## Save results to  ./data-unshared/derived/0-greeted.rds
```

```
saveRDS(ds1, path_save)
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()
```

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows >= 8 x64 (build 9200)
```

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

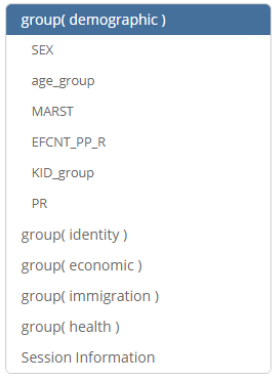
- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of linked dynamic document



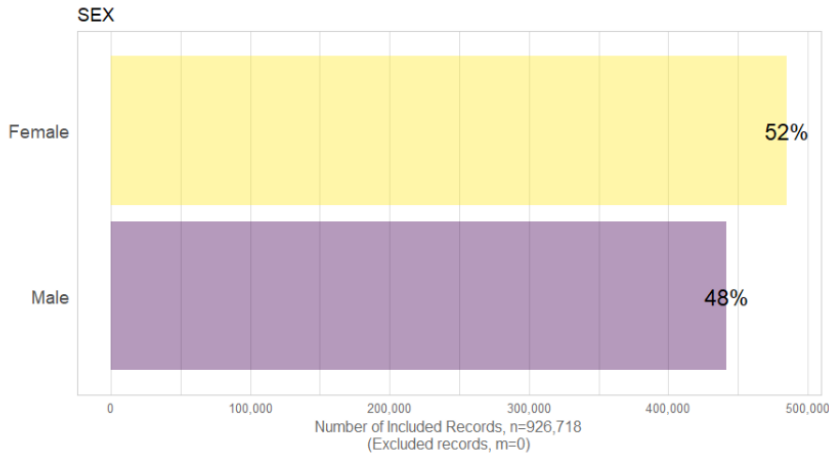

```
$ KID_group          <fct> one or two children, three or more children, no children, one or two...
$ YRIM_group         <fct> 2002 or later, 2002 or later, Non-immigrants and institutional resid...
$ age_group          <fct> 40 to 44, 30 to 34, 65 to 69, 19 to 24, 55 to 59, 70 to 74, 30 to 34...
```

This chunk will subset the data

```
# this chunk is called by ./reports/eda-1/eda-1a-first-gen-immigrant.Rmd
ds <- ds %>%
  # dplyr::filter(PR %in% selected_provinces) %>%
  dplyr::filter(IMMDER   == "Immigrants") %>%
  dplyr::filter(GENSTPOB == "1st generation - Respondent born outside Canada")
```

group(demographic)

SEX



$SEX $SEX levels 1 2 "Female" "Male"

$SEX $SEX label [1] "Sex"

$SEX $SEX description [1] "Sex"

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

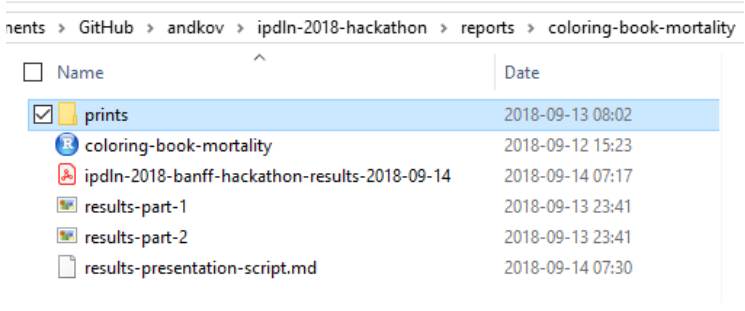
- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

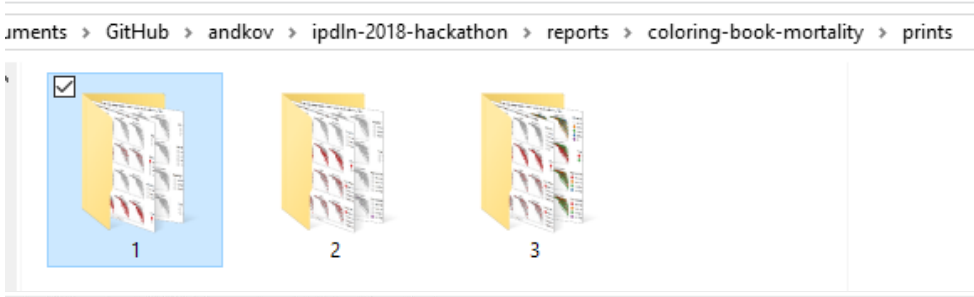
Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

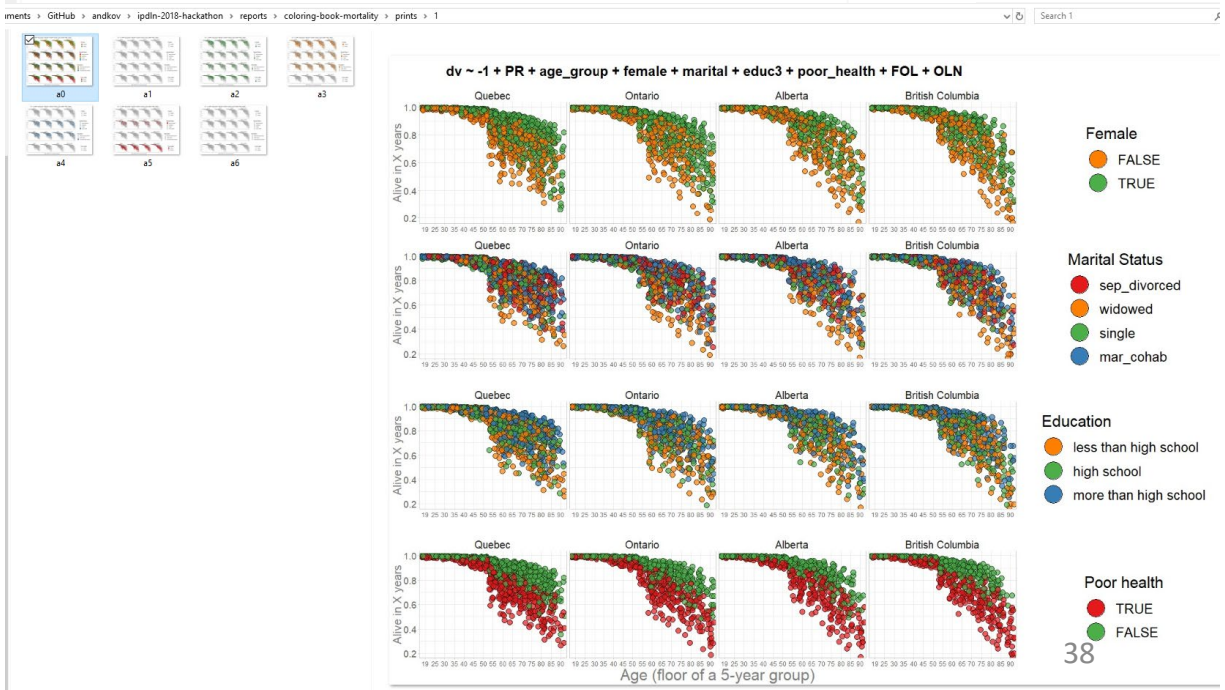
- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of project repository





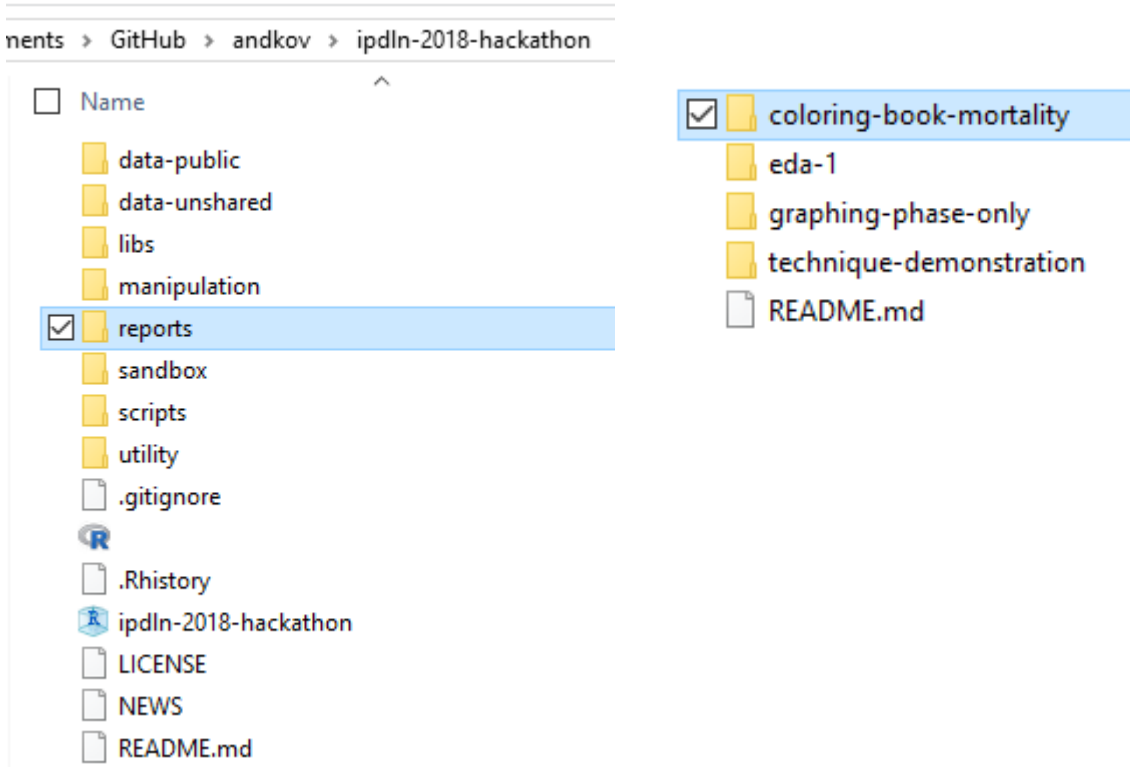


B. Workflow Highlights

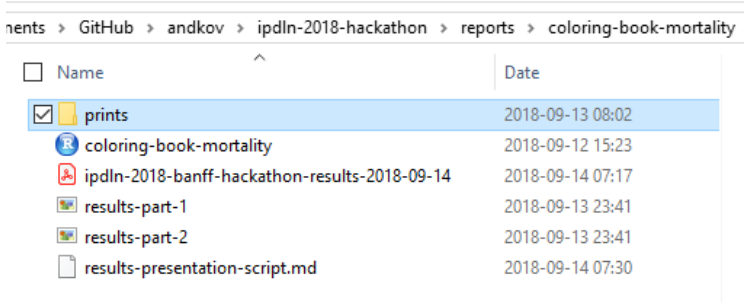
1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

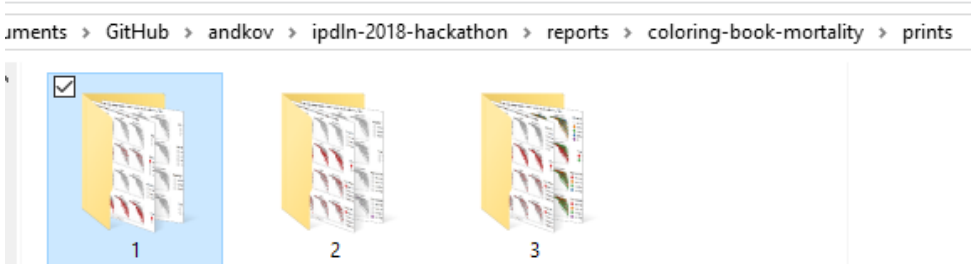
./reports/coloring-book-mortality/

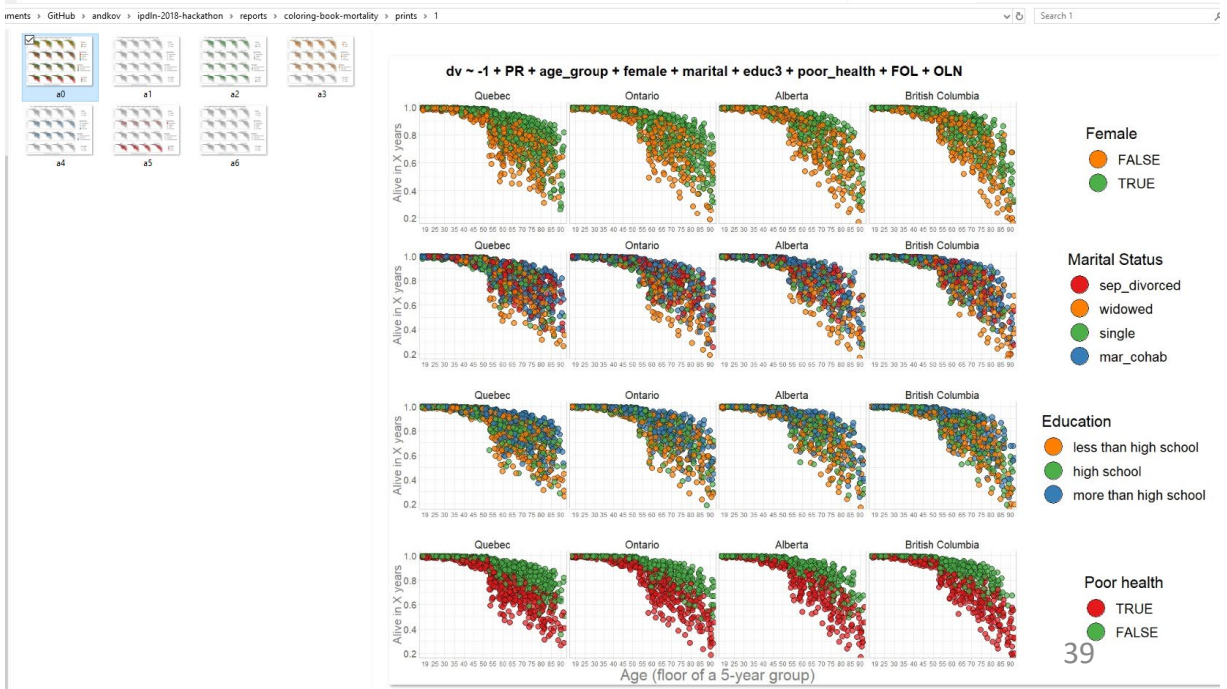
Fails to separate modeling, graphing, and reporting



Screenshots of project repository







B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

Branch: master ▾ ipdln-2018-hackathon / README.md

andkov Update README.md

Technique demonstration

- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of coloring-book-mortality.R script. Optimized for learning the workflow with the original data. For full details consult its stitched_output.
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during technical-demonstration, stored in `./data-public/dereived/technique-demonstration/`. For full details consult its stitched_output

nents › GitHub › andkov › ipdln-2018-hackathon

Name
data-public
data-unshared
libs
manipulation
reports
sandbox
scripts
utility
.gitignore
.Rhistory
ipdln-2018-hackathon
LICENSE
NEWS
README.md

coloring-book-mortality
eda-1
graphing-phase-only
technique-demonstration
README.md

ents › GitHub › andkov › ipdln-2018-hackathon › reports › graphing-phase-only

Name	Date modified	Type	Size
figure-png	2018-10-30 12:27	File folder	
prints	2018-10-30 12:58	File folder	
stitched_output	2018-10-30 13:48	File folder	
graphing-phase-only.md	2018-10-30 13:40	MD File	24 KB
graphing-phase-only	2018-10-30 13:43	R File	16 KB
graphing-phase-only	2018-10-30 13:36	RMD File	5 KB
graphing-phase-only-1	2018-10-30 13:37	Chrome HTML Do...	2,805 KB
graphing-phase-only-2	2018-10-30 13:40	Chrome HTML Do...	2,771 KB

B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

.R – stores analysis (what really happens)

graphing-phase-only.R | graphing-phase-only.Rmd

Source on Save | Run | Source

```
# Run next lint to stitch a tech report of this script (used only in RStud
# knitr::stitch_rmd( script = "./reports/graphing-phase-only/graphing-phas

rm(list=ls(all=TRUE)) #Clear the memory of variables from previous run.
# This is not called by knitr, because it's above the first chunk.
cat("\f") # clear console when working in RStudio

# ---- load-sources ------------------------------------------------------
# Call `base::source()` on any repo file that defines functions needed bel
# Ideally, no real operations are performed.
base::source("./scripts/graphing/graph-logistic.R")
base::source("./scripts/graphing/graph-presets.R") # fonts, colors, themes
# ---- load-packages -----------------------------------------------------
# Attach these packages so their functions don't need to be qualified: htt
library(ggplot2) #For graphing
library(magrittr) # Pipes
library(dplyr)
requireNamespace("dplyr", quietly=TRUE)
requireNamespace("TabularManifest") # devtools::install_github("Melinae/Ta
requireNamespace("knitr")
requireNamespace("scales") #For formating values in graphs
requireNamespace("RColorBrewer")

# ---- declare-globals ---------------------------------------------------
# link to the source of the location mapping

# This script works with model results data estimated during /technique-de
path_input_micro <- "./data-public/derived/technique-demonstration/ls_mode
path_input_meta  <- "./data-unshared/derived/ls_guide.rds"

# test whether the file exists / the link is good
testit::assert("File does not exist", base::file.exists(path_input_micro))
testit::assert("File does not exist", base::file.exists(path_input_meta))

# declare where you will store the product of this script
```

load-sources
load-packages
declare-globals
load-data
tweak-data
inspect-data
define-utility-functi...
quick_save
define-graph-controls
define-coloring-bo...
assign_color
print-display-0
print-display-1
print-display-2
print-display-3
print-display-4
print-display-5
print-display-6
save-to-disk
publish

295:12 | print-display-4 | R Script

.Rmd – stores presentation (how you tell about it)

graphing-phase-only.R | graphing-phase-only.Rmd

Knit | Insert | Run

```
---
title: "Graphing Phase Only"
date: "Date: `r Sys.Date()`"
output:
  html_document:
    keep_md: yes
    toc: yes
    toc_float: yes
    output:
    highlight: tango
    theme: spacelab
---
This report provides demonstration of visualizing the model estimated by
`./reports/technique-demonstration/technique-demonstration.R`
It does not require access to the raw data, instead it loads only
estimated model results.


<!--  Set the working directory to the repository's base directory; this
assumes the report is nested inside of two directories.-->
```{r, echo=F, message=F}
# cat("Working directory: ", getwd())
library(knitr)
opts_knit$set(root.dir='../../')  #Don't combine this call with any other
chunk -especially one that uses file paths.
```


<!-- Set the report-wide options, and point to the external code file.
-->
```{r set-options, echo=F}
# cat("Working directory: ", getwd())
report_render_start_time <- Sys.time()
opts_chunk$set(
  results      = 'show',
  comment      = NA,
  tidy         = FALSE,
  # dpi          = 400,
  out.width  = "900px", #This affects only the markdown, not the
underlying png file.  The height will be scaled appropriately.
  fig.width    = 12,
  fig.height   = 6,
  fig.path     = 'figure-png/'
)

echo_chunks    <- TRUE #Toggle for debugging.
message_chunks <- FALSE #Toggle for debugging.
options(width=100) #So the output is 25% wider than the default.
read_chunk("./reports/graphing-phase-only/graphing-phase-only.R") #This
allows knitr to call chunks tagged in the underlying *.R file.
```
```

load sources
load packages
declare globals
load data
tweek data
inspect data
custom functions
Graph controls
Coloring Book setting
Display 0
Display 1
Display 2
Display 3
Display 4
Display 5
Display 6
save to disk
Session Information

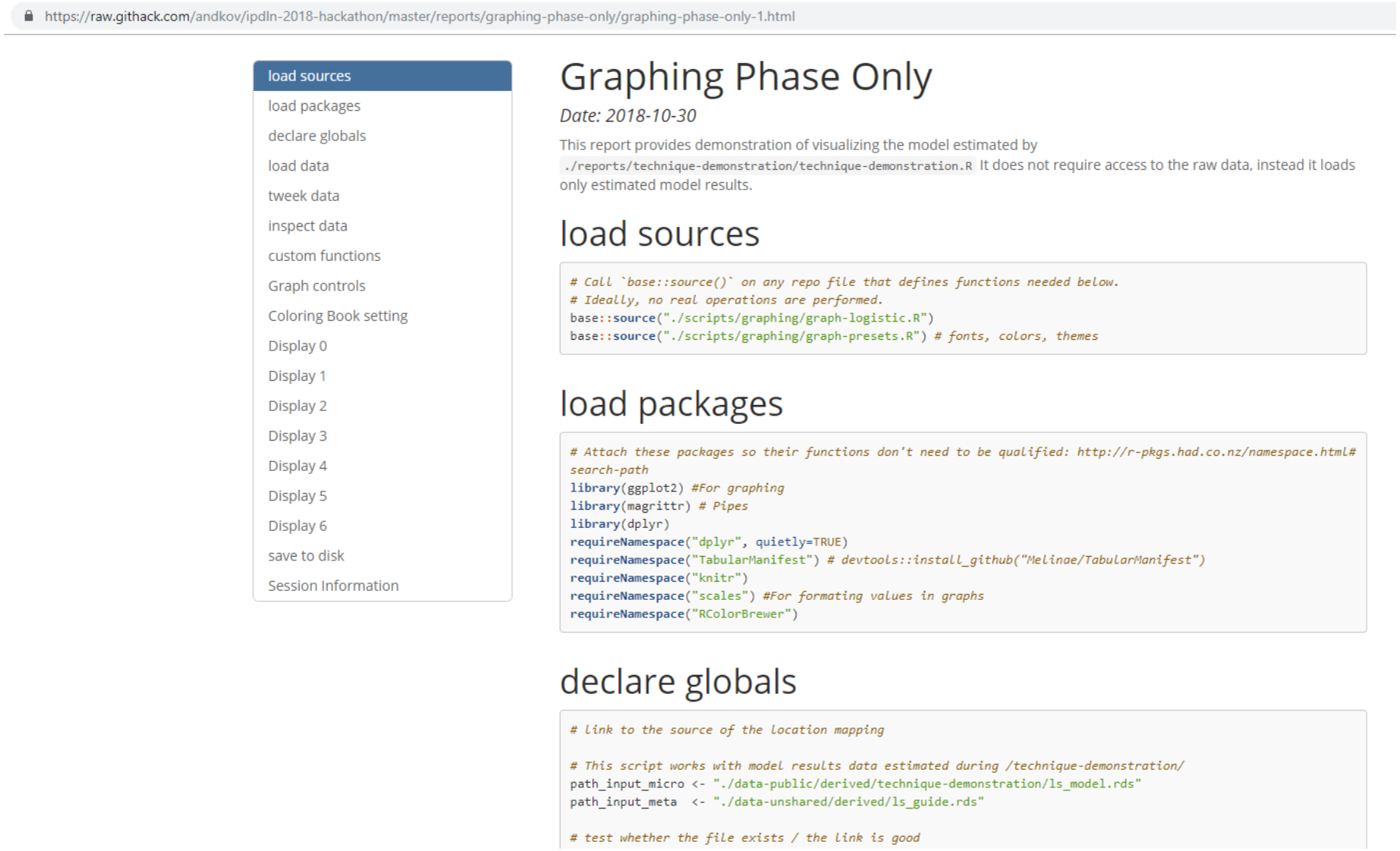
R Markdown

B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

.R – stores analysis (what really happens)

.Rmd – stores presentation (how you tell about it)



B. Workflow Highlights

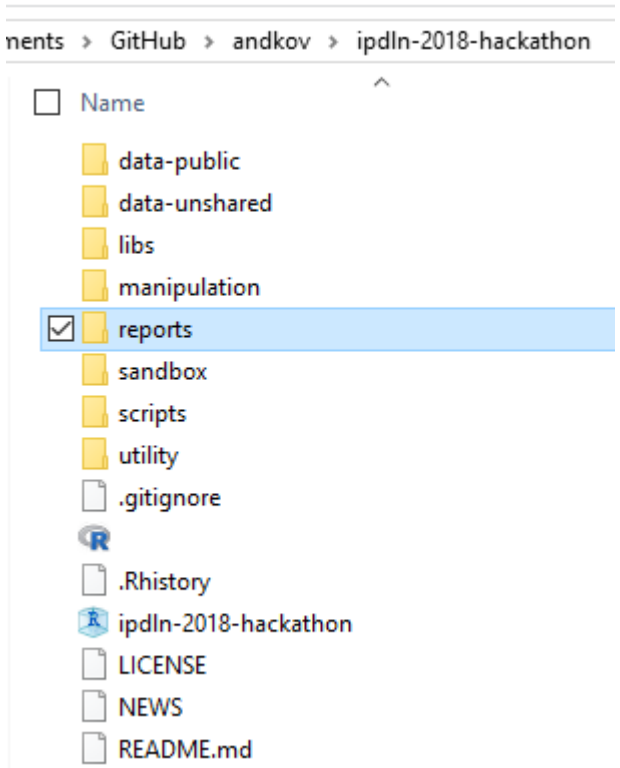
1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

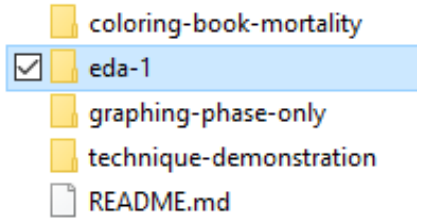
Branch: master ▾ ipdln-2018-hackathon / README.md

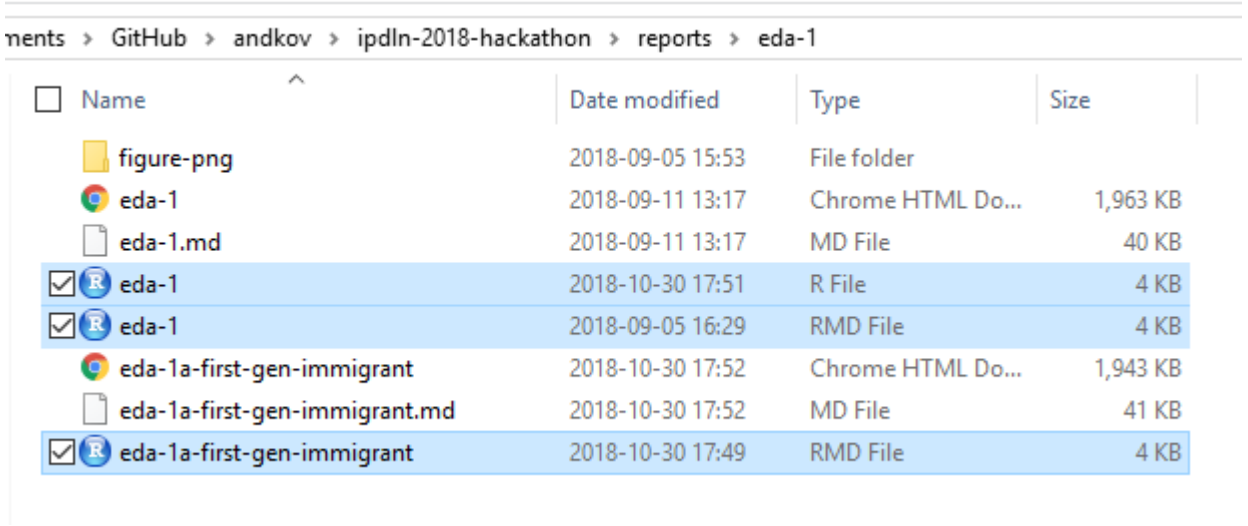
andkov Update README.md

Technique demonstration

- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of coloring-book-mortality.R script. Optimized for learning the workflow with the original data. For full details consult its stitched_output.
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during technical-demonstration, stored in `./data-public/dereived/technique-demonstration/`. For full details consult its stitched_output







B. Workflow Highlights

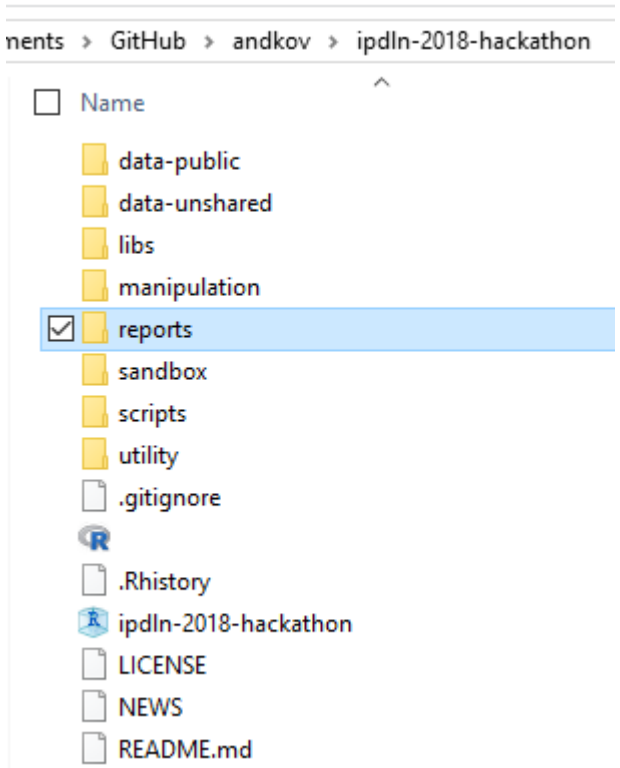
1.4 Two essential means of production: knitr::stitch() vs rmarkdown::render()

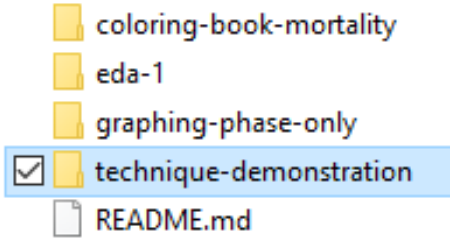
Branch: master ▾ ipdln-2018-hackathon / README.md

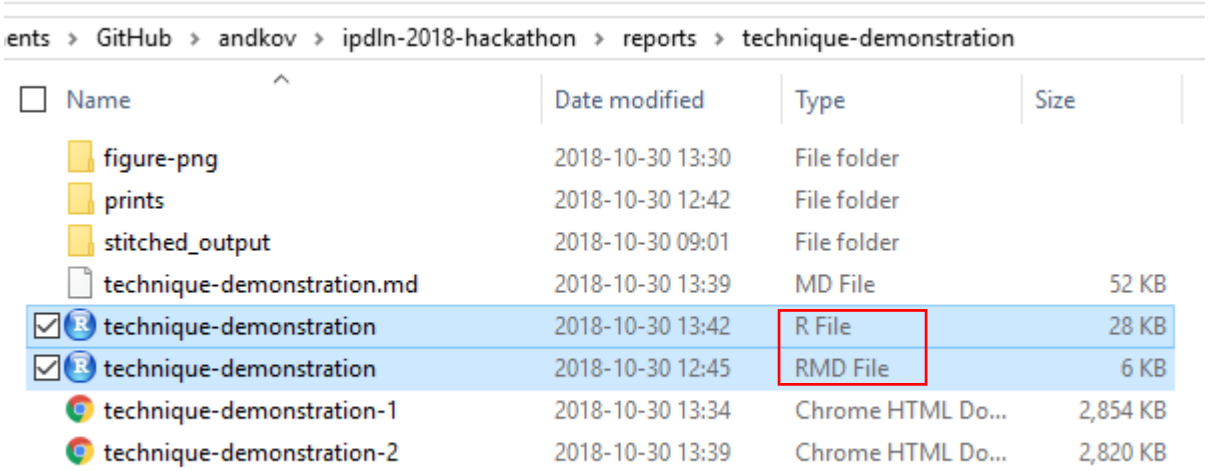
andkov Update README.md

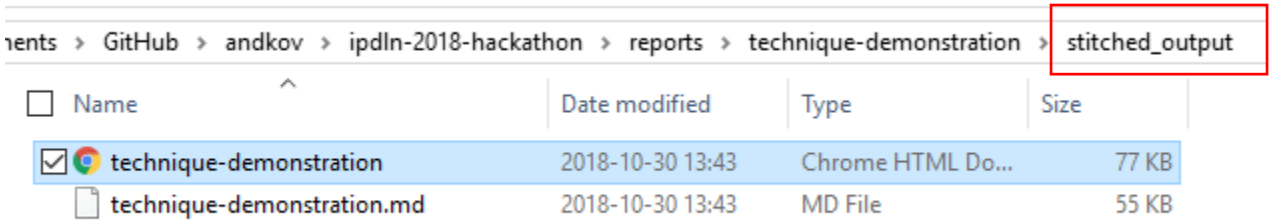
Technique demonstration

- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of coloring-book-mortality.R script. Optimized for learning the workflow with the original data. For full details consult its stitched_output.
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during technical-demonstration, stored in `./data-public/dereived/technique-demonstration/`. For full details consult its stitched_output









A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff
0.1 Modeling form: univariate logistic regression with categorical predictors
0.2 Graphical form: faceted scatterplot in ggplot2
0.3 Coloring book: Mapping informed expectations from predictors onto color

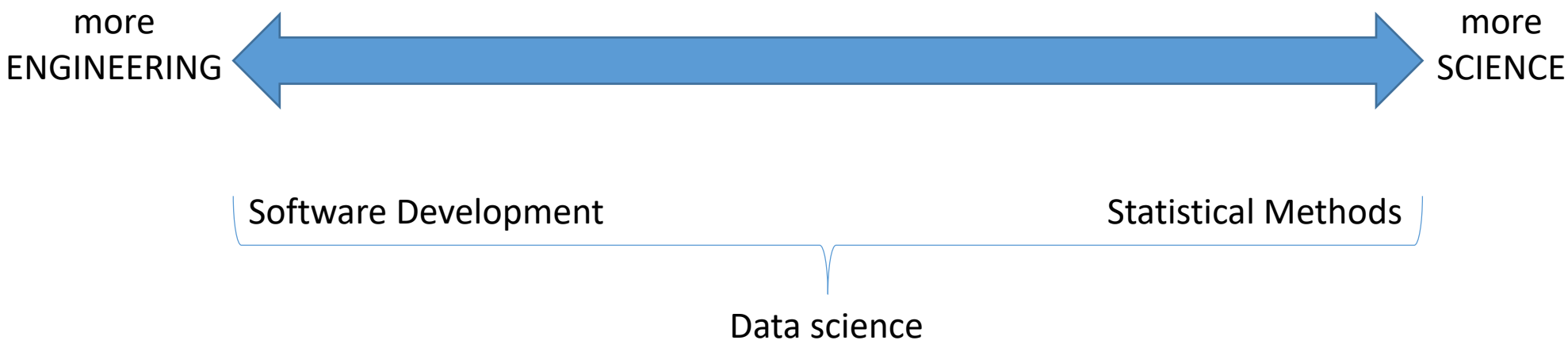
B. Workflow Highlights

1.0 "Let no one ignorant of geometry enter": (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Autonomous phases: data cleaning, statistical modelling, graph production
1.3 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
1.4 Two essential means of production: knitr::stitch() vs rmarkdown::render()

C. Conclusions

2.0 Different than Notebooks: sacrifices simplicity for agility via layers of isolation
2.1 R (+ .Rmd) = .html (+ .pdf) : moving away from *data playing* towards *data science*
2.2 Reproducible projects: moving away from notebooks towards software
2.3 Looking back to Neil Ernst talk:

- Parameters and configuration
- Hidden state
- Longevity and version control
- Testing and modularity
- Notebook carpentry

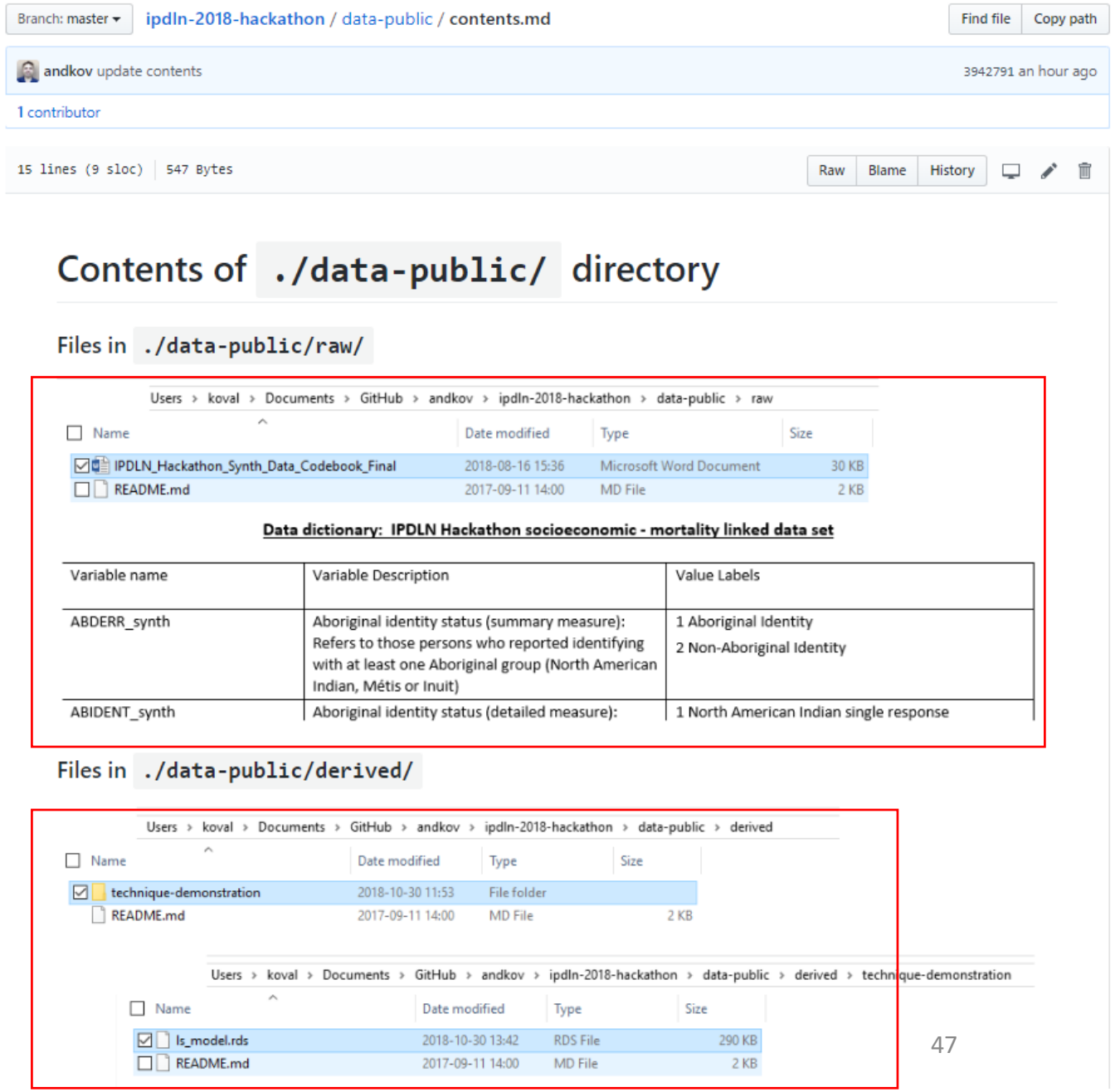

more
ENGINEERING
more
SCIENCE
Software Development
Statistical Methods
Data science

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff

Provided data codebook has been transformed into a list object `0-ls_guide.rds` By ./manipulation/0-metador.R and stored in the ./data-unshared/derived/ so that each user must re-create it from script.

./reports/technique-demonstration/ walks through application of the technique with additional annotation for learners



Branch: master ▾ ipdln-2018-hackathon / data-public / contents.md

Find file | Copy path

andkov update contents 3942791 an hour ago

1 contributor

15 lines (9 sloc) | 547 Bytes

Raw | Blame | History

Contents of ./data-public/ directory

Files in ./data-public/raw/

Users › koval › Documents › GitHub › andkov › ipdln-2018-hackathon › data-public › raw

Name	Date modified	Type	Size
IPDLN_Hackathon_Synth_Data_Codebook_Final	2018-08-16 15:36	Microsoft Word Document	30 KB
README.md	2017-09-11 14:00	MD File	2 KB

Data dictionary: IPDLN Hackathon socioeconomic - mortality linked data set

Variable name	Variable Description	Value Labels
ABDERR_synth	Aboriginal identity status (summary measure): Refers to those persons who reported identifying with at least one Aboriginal group (North American Indian, Métis or Inuit)	1 Aboriginal Identity 2 Non-Aboriginal Identity
ABIDENT_synth	Aboriginal identity status (detailed measure):	1 North American Indian single response

Files in ./data-public/derived/

Users › koval › Documents › GitHub › andkov › ipdln-2018-hackathon › data-public › derived

Name	Date modified	Type	Size
technique-demonstration	2018-10-30 11:53	File folder	
README.md	2017-09-11 14:00	MD File	2 KB

Users › koval › Documents › GitHub › andkov › ipdln-2018-hackathon › data-public › derived › technique-demonstration

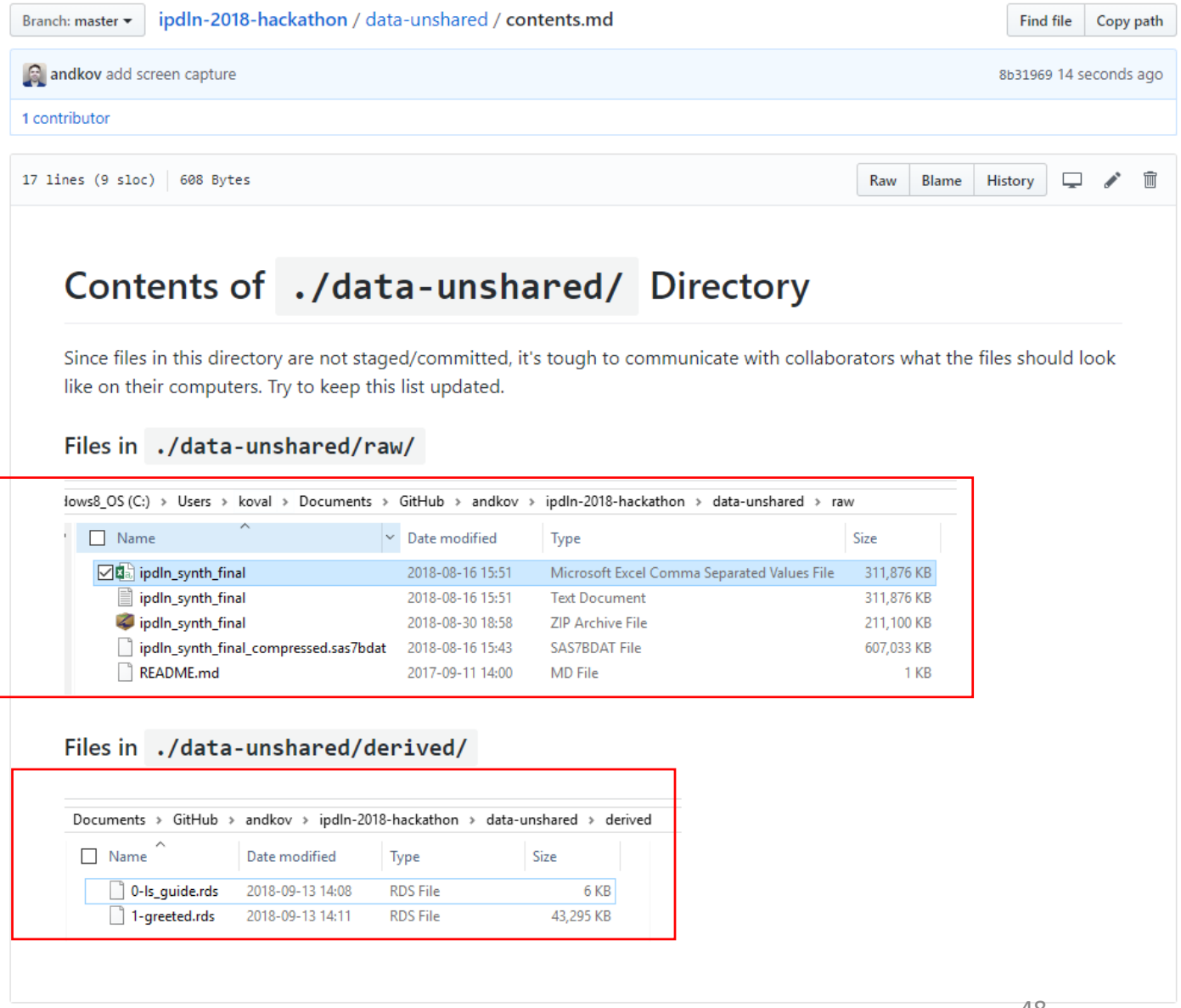
Name	Date modified	Type	Size
ls_model.rds	2018-10-30 13:42	RDS File	290 KB
README.md	2017-09-11 14:00	MD File	2 KB

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff

Participants received a package raw data files and data codebook

./manipulation/1-greeter.R prepares raw data for exploration and generic modeling



```r
# basic inspection
ds0 %>% dplyr::glimpse(50)
```

```
## Observations: 4,346,649
## Variables: 34
## $ ABDERR_synth              <int> 2, 2, 2, 2, ...
## $ ABIDENT_synth             <int> 6, 6, 6, 6, ...
## $ ADIFCLTY_synth            <int> 1, 1, 1, 1, ...
## $ CITSM_synth               <int> 2, 2, 1, 1, ...
## $ COWD_synth                <int> 4, 4, 7, 4, ...
## $ DISABFL_synth             <int> 1, 1, 4, 1, ...
## $ DISABIL_synth             <int> 9, 9, 14, 9,...
## $ DVISMIN_synth             <int> 14, 14, 14, ...
## $ FOL_synth                 <int> 1, 1, 2, 1, ...
## $ FPTIM_synth               <int> 1, 1, 3, 2, ...
## $ GENSTPOB_synth            <int> 1, 1, 3, 3, ...
## $ HCDD_synth                <int> 9, 8, 1, 2, ...
## $ IMMDER_synth              <int> 1, 1, 3, 3, ...
## $ LOINCA_synth              <int> 1, 1, 1, 1, ...
## $ LOINCB_synth              <int> 1, 1, 1, 2, ...
## $ MARST_synth               <int> 2, 2, 2, 4, ...
## $ NOCSBRD_synth             <int> 4, 4, 11, 6,...
## $ OLN_synth                 <int> 3, 1, 2, 3, ...
## $ POBDER_synth              <int> 3, 3, 1, 1, ...
## $ SEX_synth                 <int> 1, 1, 1, 1, ...
## $ TRMODE_synth              <int> 2, 2, 9, 5, ...
## $ RPAIR_synth               <int> 3, 1, 1, 2, ...
## $ PR_synth                  <int> 35, 46, 24, ...
## $ RUINDFG_synth             <int> 1, 1, 2, 2, ...
## $ d_licoratio_da_bef_synth <int> 5, 3, 3, 2, ...
## $ S_DEAD_synth              <int> 2, 2, 1, 2, ...
## $ EFCNT_PP_R_synth          <int> 4, 5, 2, 4, ...
## $ AGE_IMM_R_group_synth     <int> 8, 6, 15, 15...
## $ COD1_synth                <int> 5, 5, 2, 5, ...
## $ COD2_synth                <int> 14, 14, 13, ...
## $ DPOB11N_synth             <int> 4, 2, 1, 1, ...
## $ KID_group_synth           <int> 2, 3, 1, 2, ...
## $ YRIM_group_synth          <int> 1, 1, 6, 6, ...
## $ age_group_synth           <int> 5, 3, 10, 1,...
```

```r
ls_model$predicted_values %>% glimpse(50) # predicted values
```

```
Observations: 3,883
Variables: 9
$ PR          <fct> Alberta, Alberta, Alberta...
$ age_group   <fct> 65, 60, 30, 80, 55, 40, 6...
$ female      <fct> FALSE, FALSE, TRUE, FALSE...
$ educ3       <fct> high school, more than hi...
$ marital     <fct> mar_cohab, mar_cohab, mar...
$ poor_health <fct> FALSE, FALSE, FALSE, TRUE...
$ FOL         <fct> English only, English onl...
$ dv_hat      <dbl> 1.8628432, 2.3139500, 6.1...
$ dv_hat_p    <dbl> 0.8656280, 0.9100258, 0.9...
```

```r
ls_model$predicted_values %>% glimpse(50) # predicted values
```

```
Observations: 3,883
Variables: 9
$ PR          <fct> Alberta, Alberta, Alberta...
$ age_group   <fct> 65, 60, 30, 80, 55, 40, 6...
$ female      <fct> FALSE, FALSE, TRUE, FALSE...
$ educ3       <fct> high school, more than hi...
$ marital     <fct> mar_cohab, mar_cohab, mar...
$ poor_health <fct> FALSE, FALSE, FALSE, TRUE...
$ FOL         <fct> English only, English onl...
$ dv_hat      <dbl> 1.8628432, 2.3139500, 6.1...
$ dv_hat_p    <dbl> 0.8656280, 0.9100258, 0.9...
```

Background – meta data

Data dictionary: IPDLN Hackathon socioeconomic - mortality linked data set

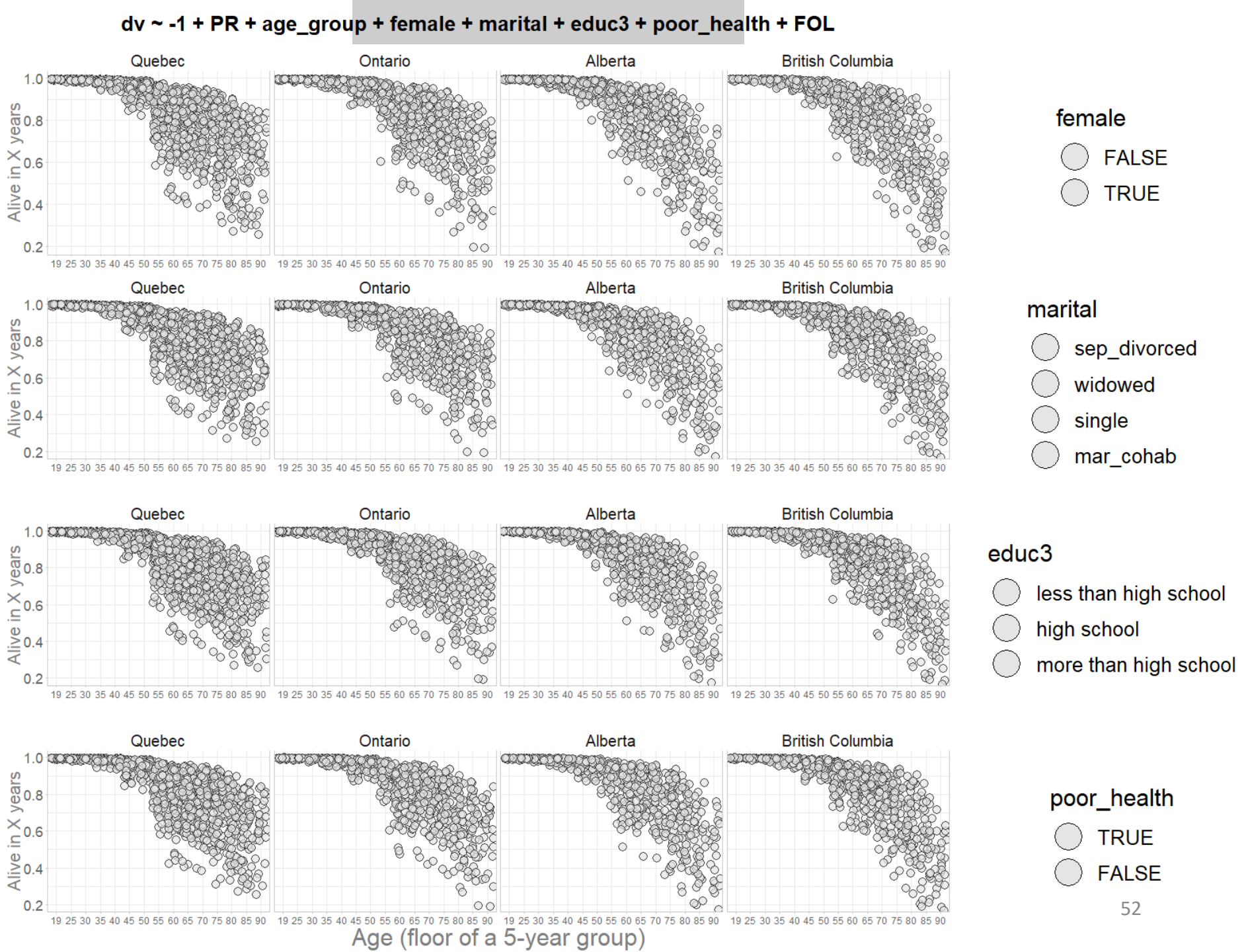
Variable name	Variable Description	Value Labels
ABDERR_synth	Aboriginal identity status (summary measure): Refers to those persons who reported identifying with at least one Aboriginal group (North American Indian, Métis or Inuit)	1 Aboriginal Identity 2 Non-Aboriginal Identity
ABIDENT_synth	Aboriginal identity status (detailed measure): Refers to those persons who reported identifying who reported identifying with at least one Aboriginal group (North American Indian, Métis or Inuit)	1 North American Indian single response 2 Métis single response 3 Inuit single response 4 Multiple Aboriginal identity responses 5 Aboriginal responses not included elsewhere 6 Non-Aboriginal identity population
ADIFCLTY_synth	Difficulties with activities of daily living: Difficulty with activities of daily living such as hearing, seeing, communicating, walking, climbing stairs, bending, learning or doing any similar activities.	1 No 2 Not stated 3 Yes, often 4 Yes, sometimes
AGE_IMM_REVISED_group_synth	Age at immigration (grouped): Refers to the age at which the respondent first obtained landed immigrant status. A landed immigrant is a person who has been granted the right to live in Canada permanently by immigration authorities.	1 < 5 years of age 2 5 to < 10 years of age 3 10 to < 15 years of age 4 15 to < 20 years of age 5 20 to < 25 years of age 6 25 to < 30 years of age 7 30 to < 35 years of age 8 35 to < 40 years of age 9 40 to < 45 years of age 10 45 to < 50 years of age 11 50 to < 55 years of age 12 55 to < 60 years of age 13 60 and over 14 Non-permanent residents 15 Non-immigrants and institutional residents
CITSM_synth	Citizenship status: Refers to the legal citizenship status of the respondent as being "Canadian citizen by birth" or something else.	1 Canadian citizen by birth 2 Not a Canadian citizen by birth
COD1_synth	Cause of death 1: Cause of death according to Global Burden of Disease Level 1 codes (with ICD-10 codes for comparison)	1 Communicable, maternal, perinatal, and nutritional conditions (GBD: U001; ICD-10: A00–B99, G00–G04, N70–N73, J00–J06, J10–J18, J20–J22, H65–H66, O00–O99, P00–P96, E00–E02, E40–E46, E50, D50–D53, D64.9, E51–64)

```
ABDERR <- list(
  "levels" = c(
    "1"  = "Aboriginal Identity"
    ,"2" = "Non-Aboriginal Identity"
    )
  ,"label" = "Aboriginal Status"
  ,"description"= "Aboriginal identity status (detailed measure): Refers to those persons who reported identif
  )
ABIDENT <- list(
  "levels" = c(
    "1"= "North American Indian single response"
    ,"2" = "Metis single response"
    ,"3" = "Inuit single response"
    ,"4" = "Multiple Aboriginal identity responses"
    ,"5" = "Aboriginal responses not included elsewhere"
    ,"6" = "Non-Aboriginal identity population"
  )
  ,"label" = "Aboriginal Identity (detail)"
  ,"description"= "Aboriginal identity status (detailed measure): Refers to those persons who reported identif
  )
ADIFCLTY <- list(
  "levels" = c(
    "1" = "No"
    ,"2" = "Not stated"
    ,"3" = "Yes, often"
    ,"4" = "Yes, sometimes"
  )
  ,"label" = "Problems with ADL"
  ,"description"= "Difficulties with activities of daily living: Difficulty with activities of daily living su
)
```

10,000 of 1st gen immigrants from each province

Informed expectation

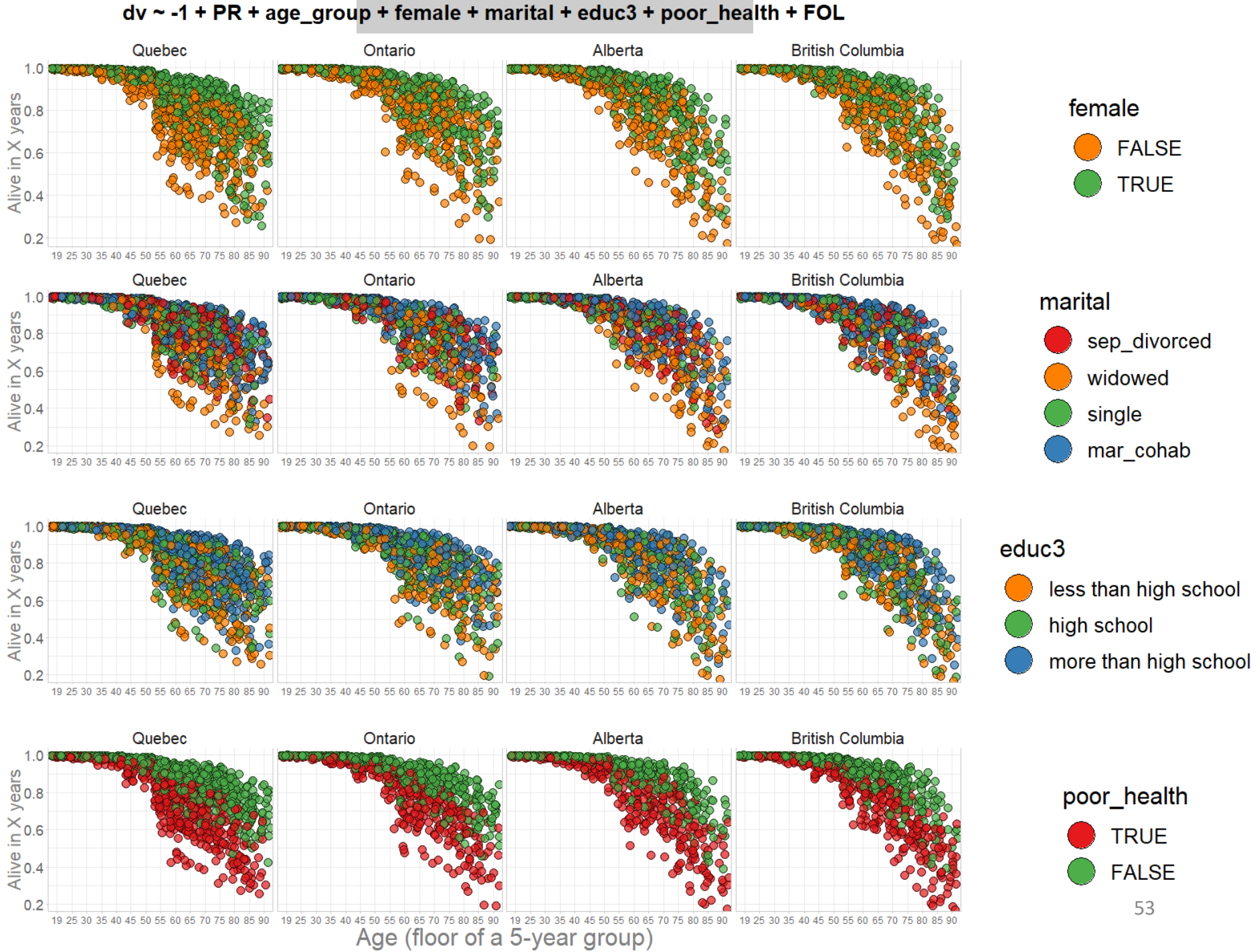
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

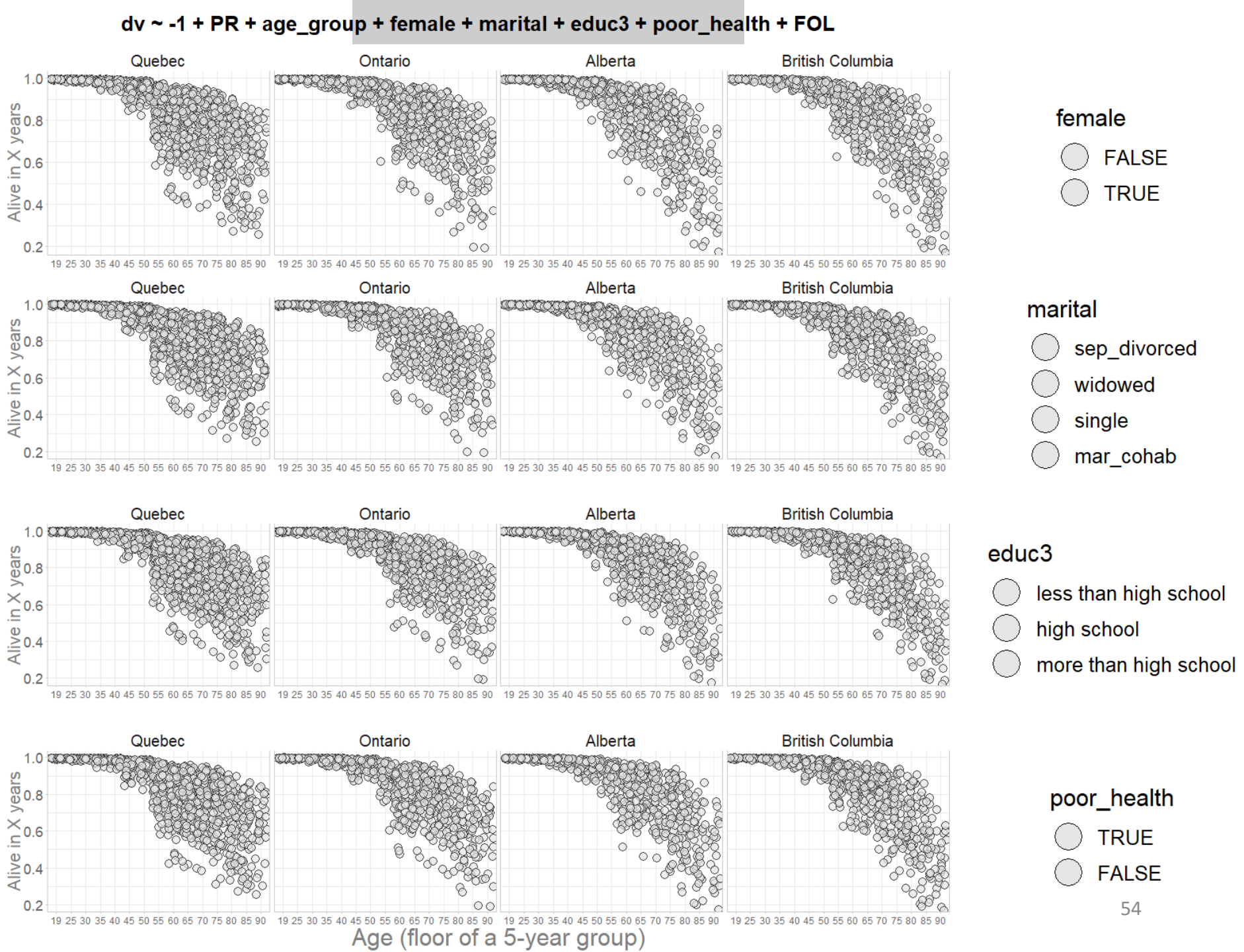
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

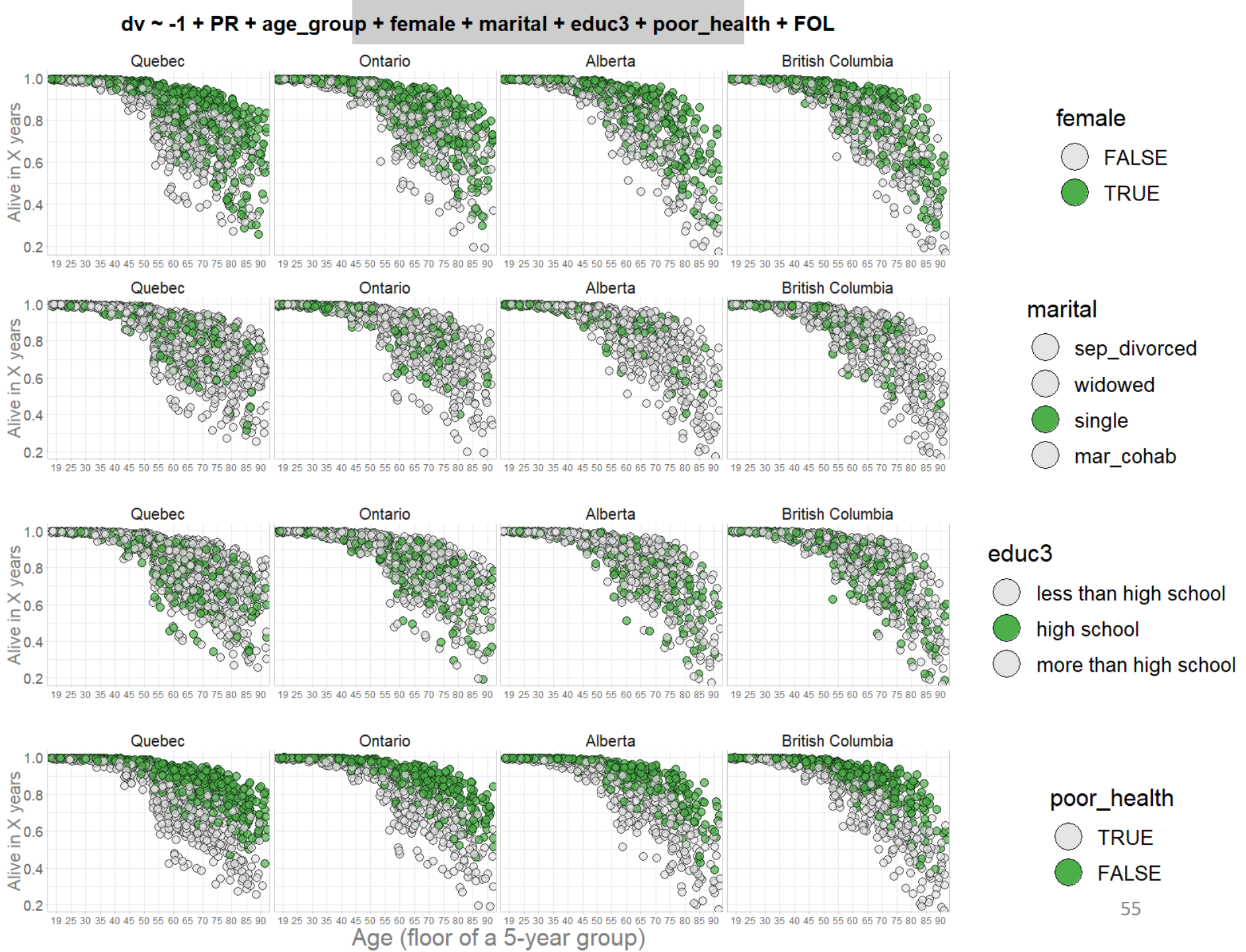
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

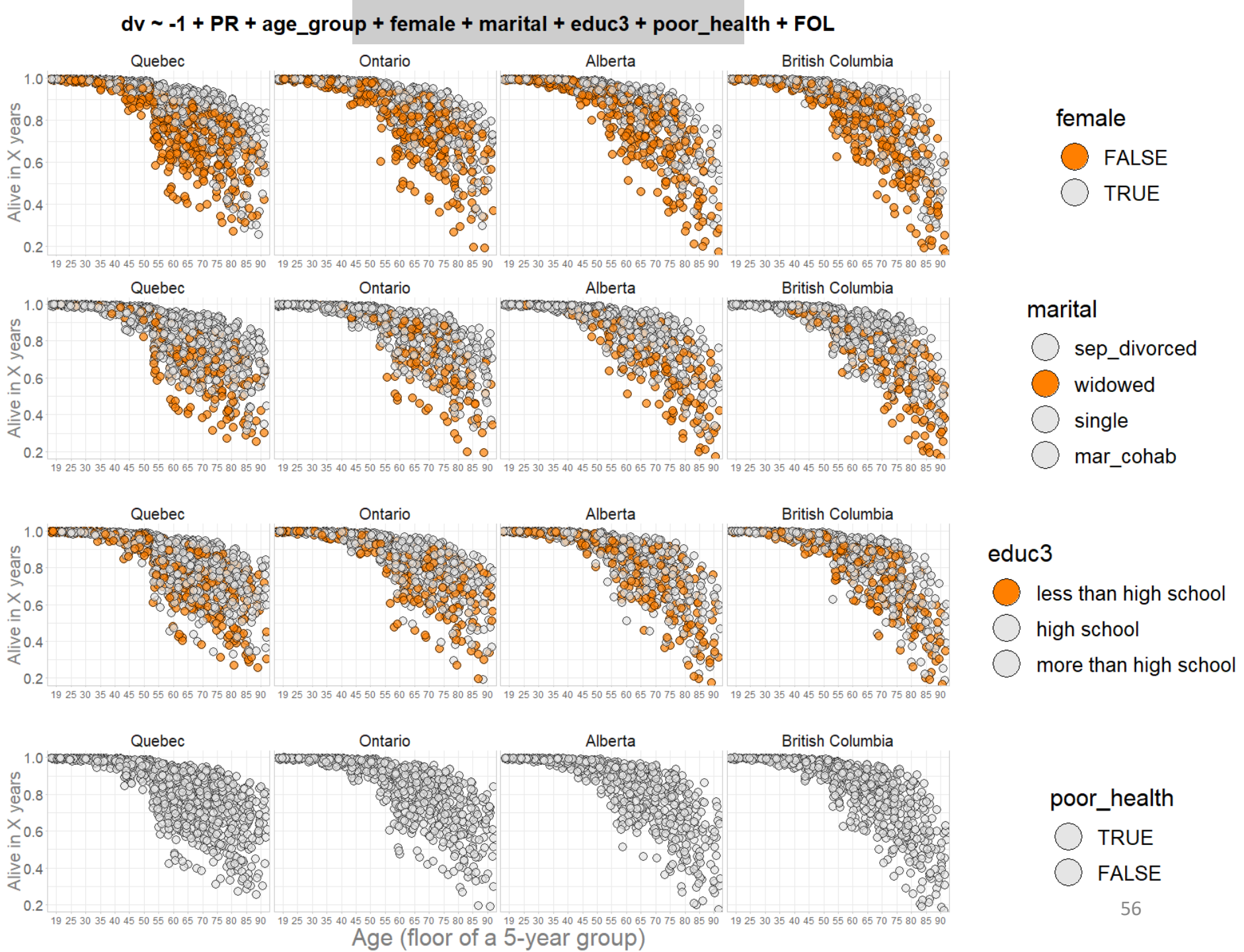
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

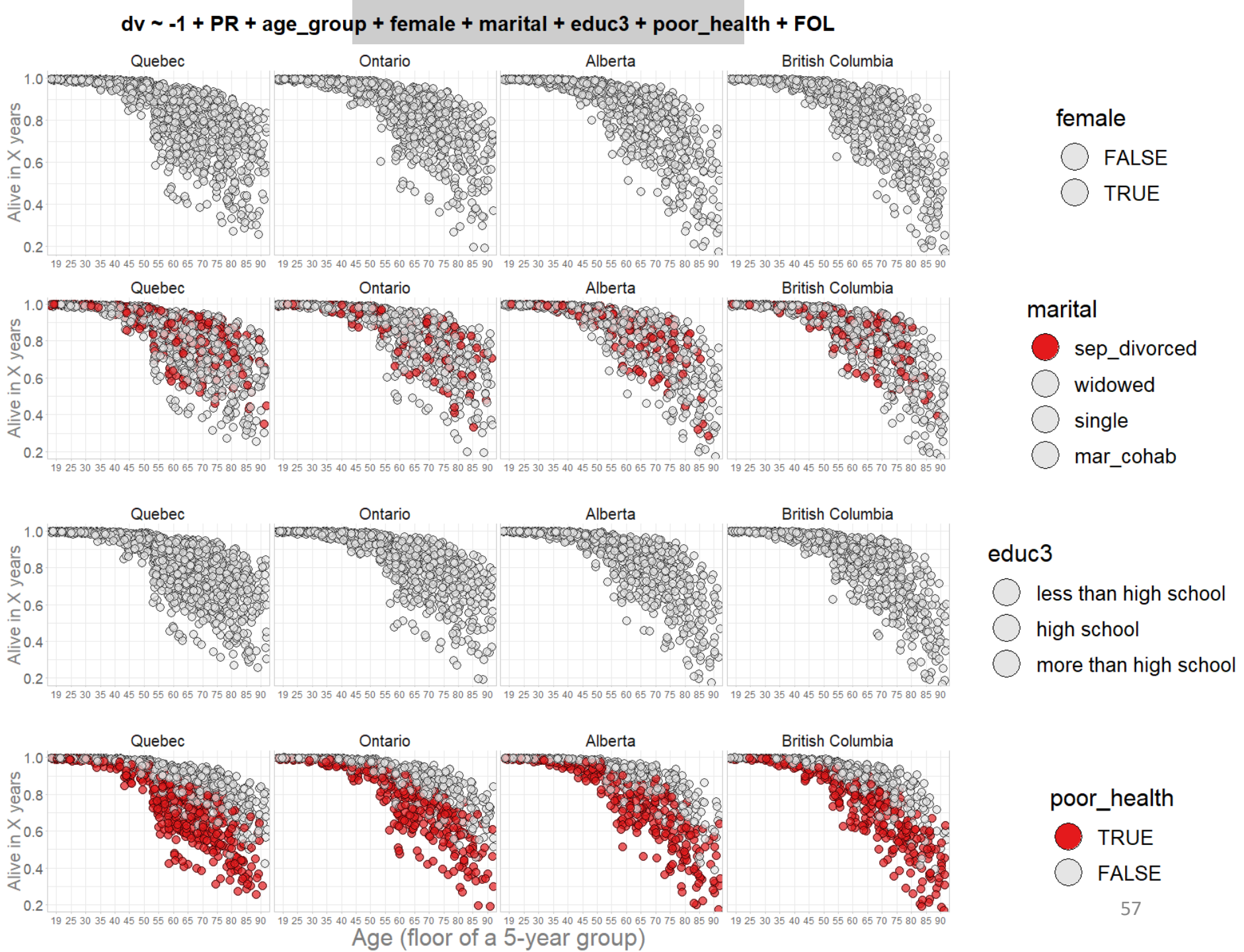
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

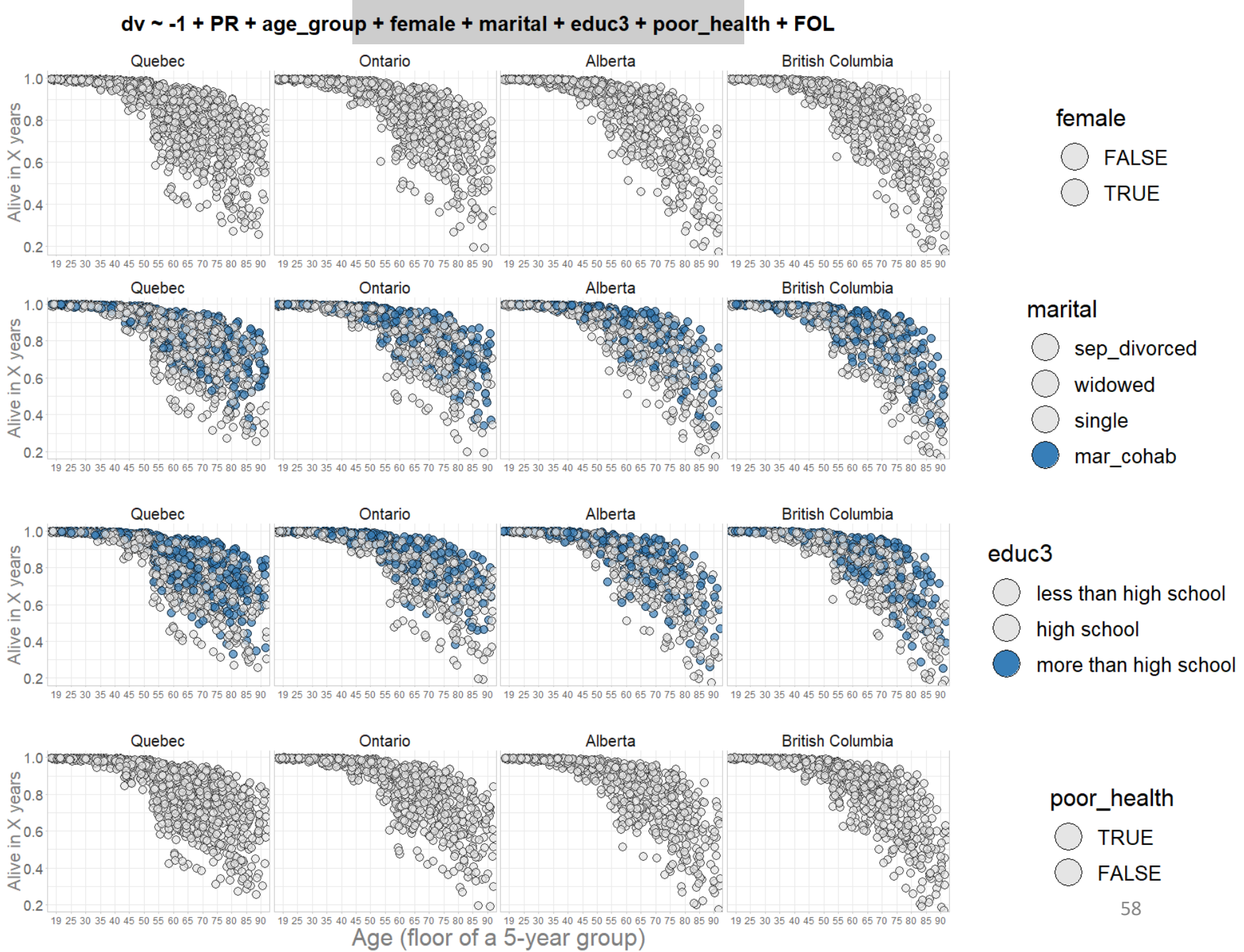
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

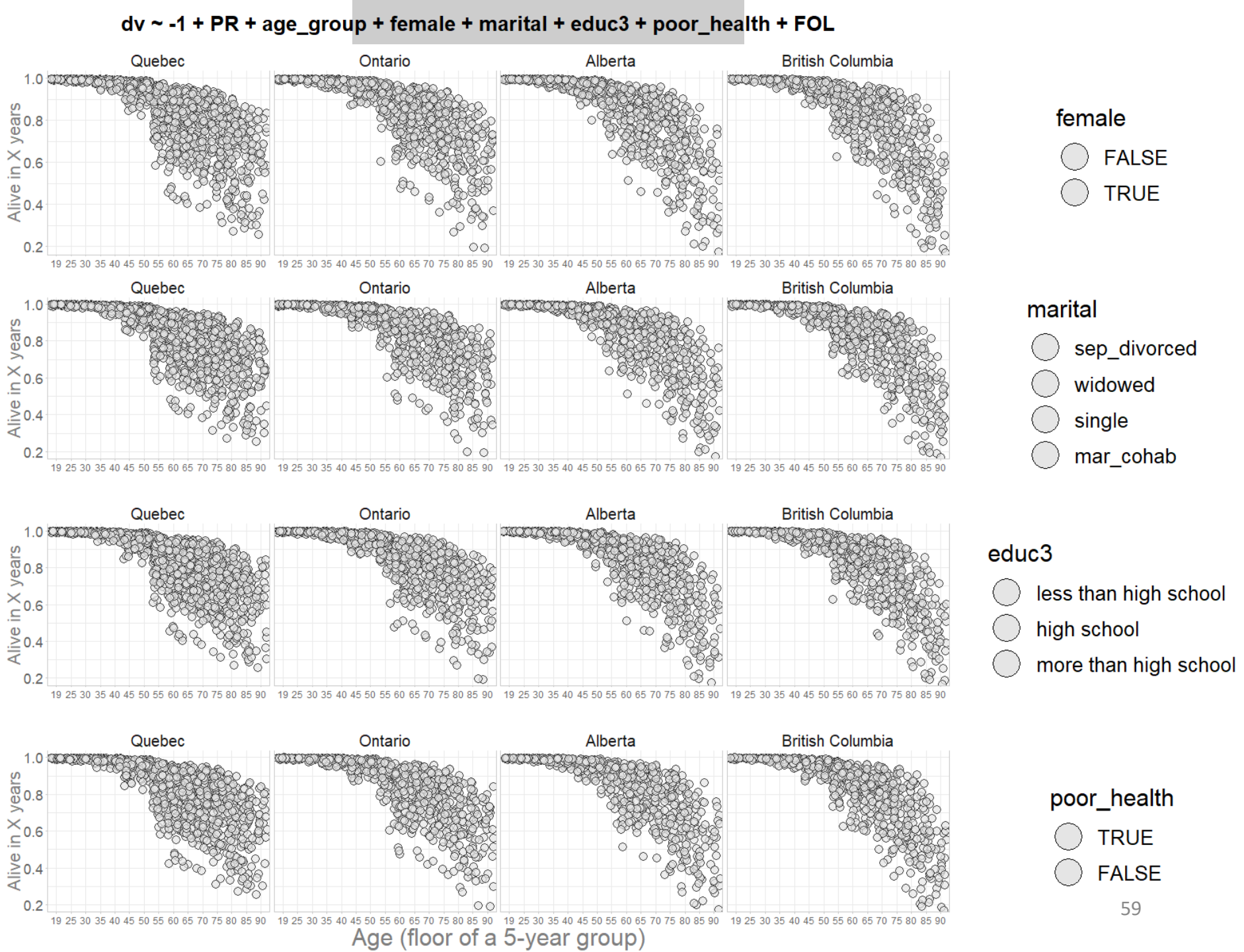
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

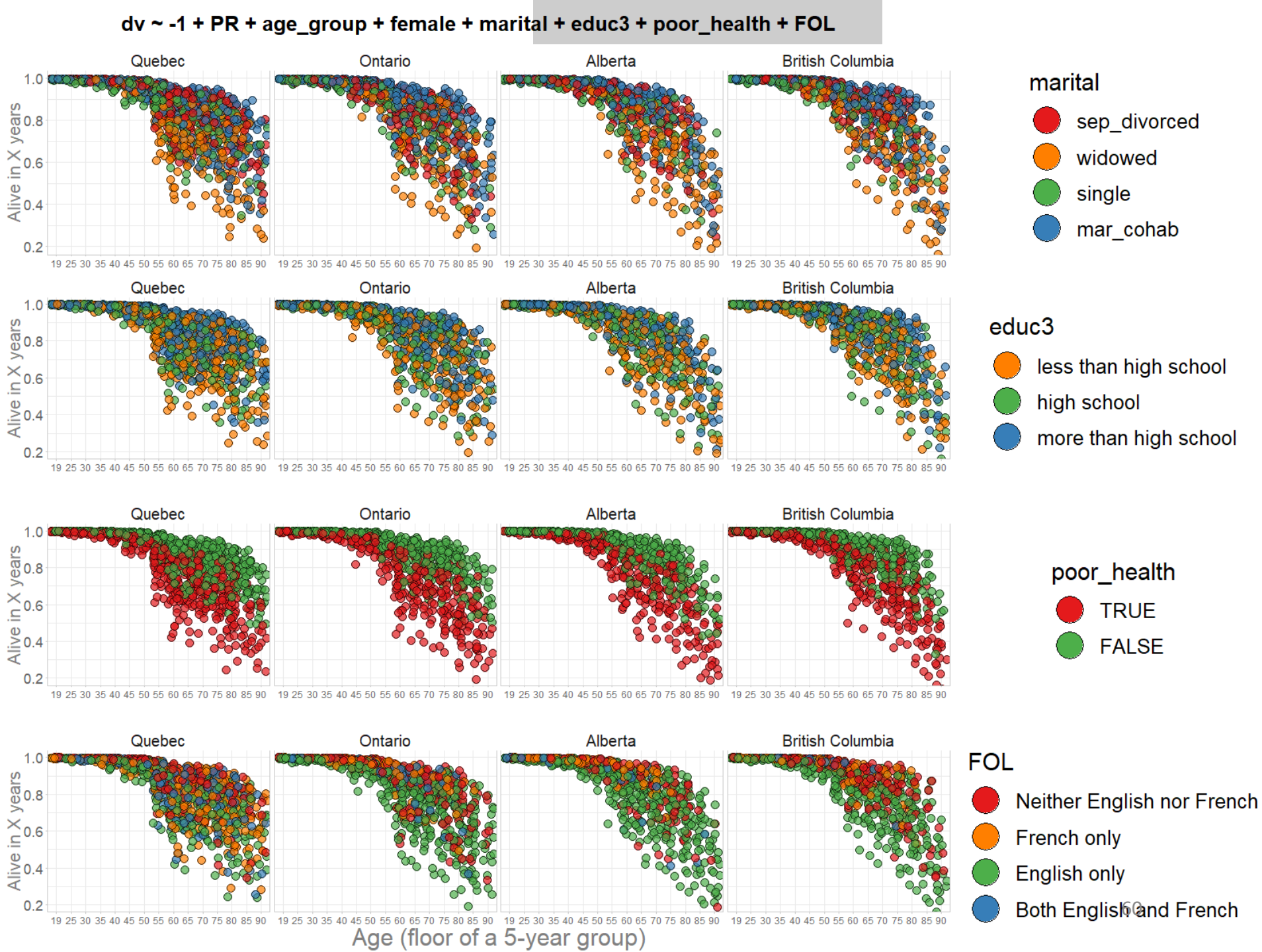
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

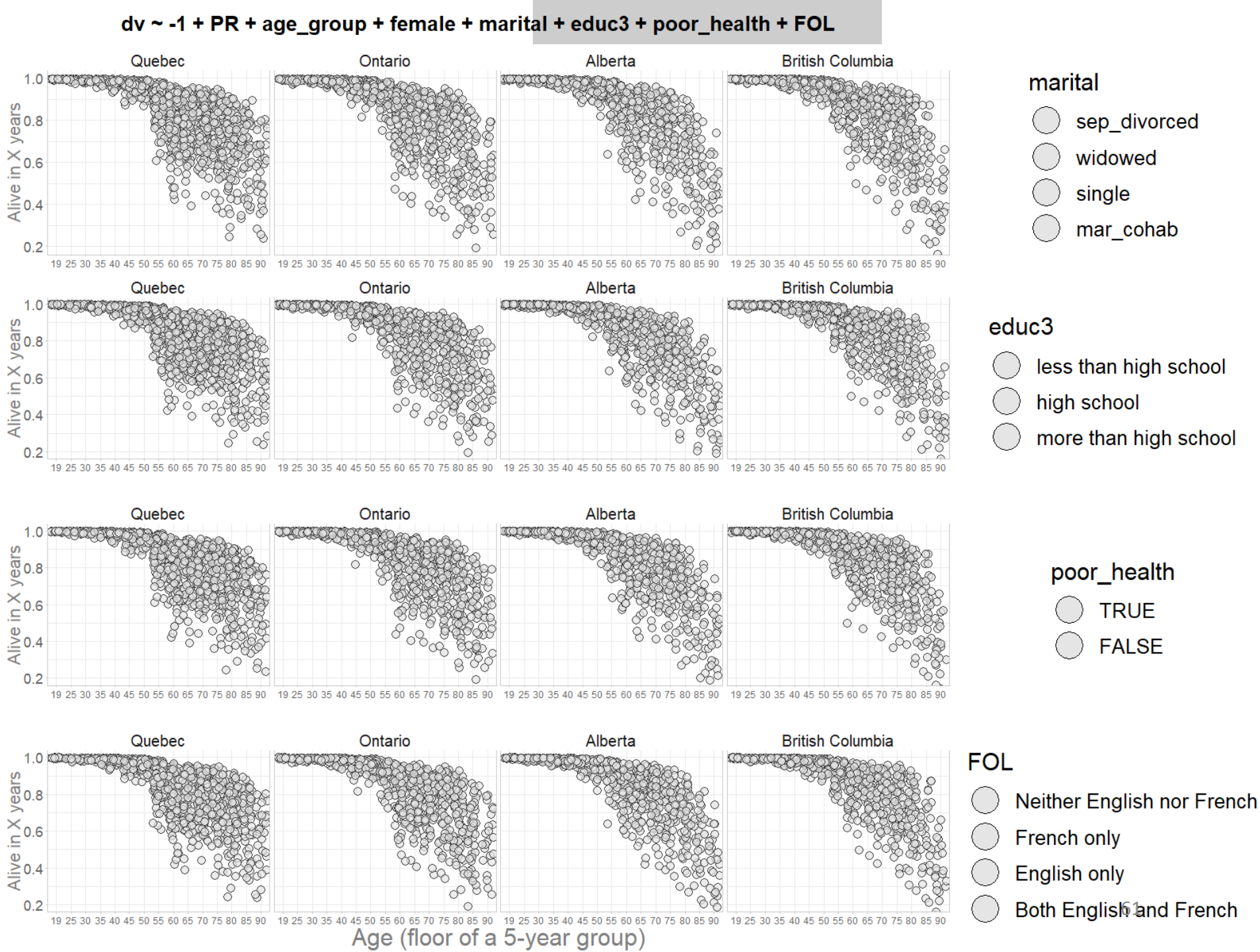
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

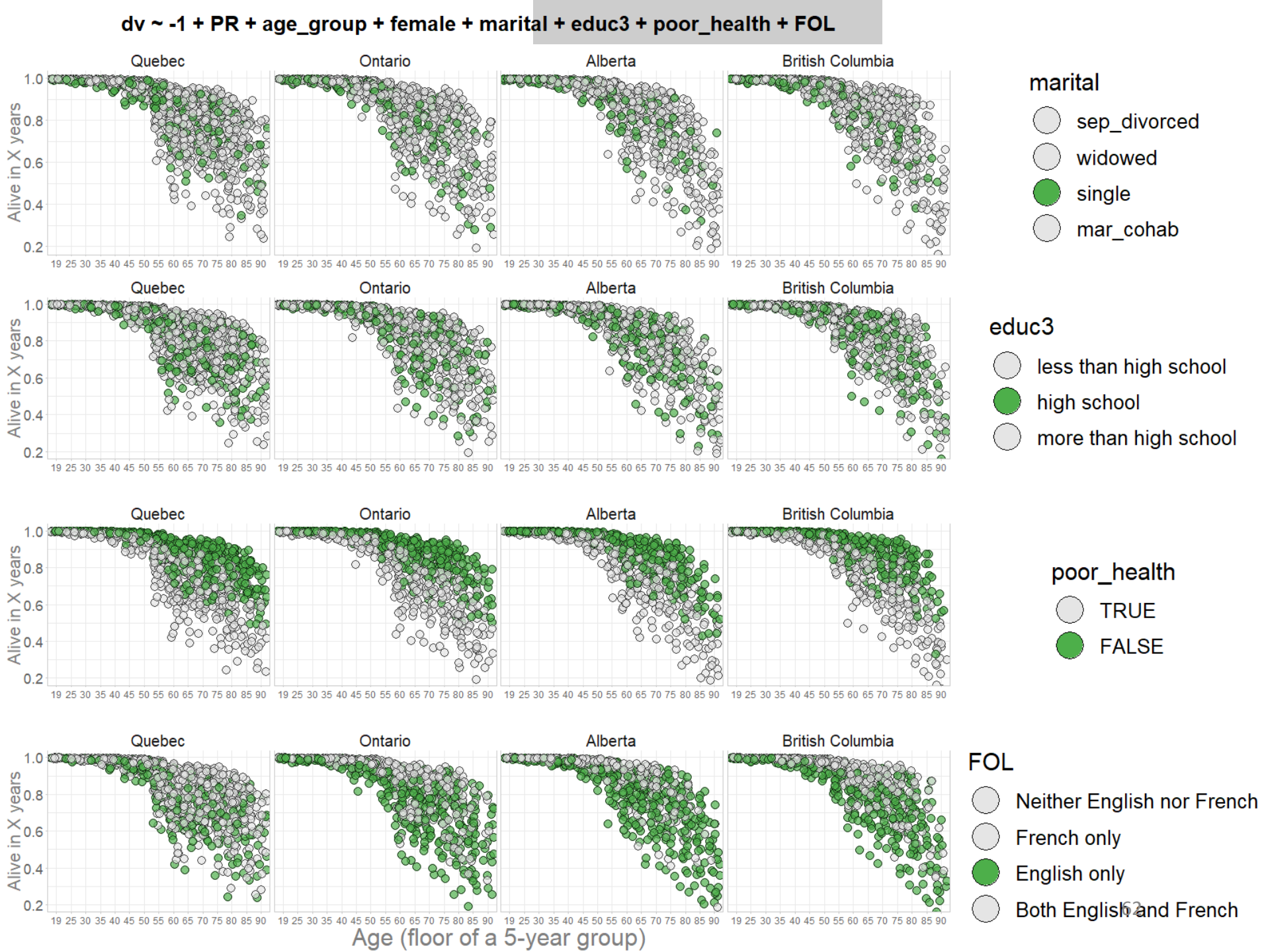
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

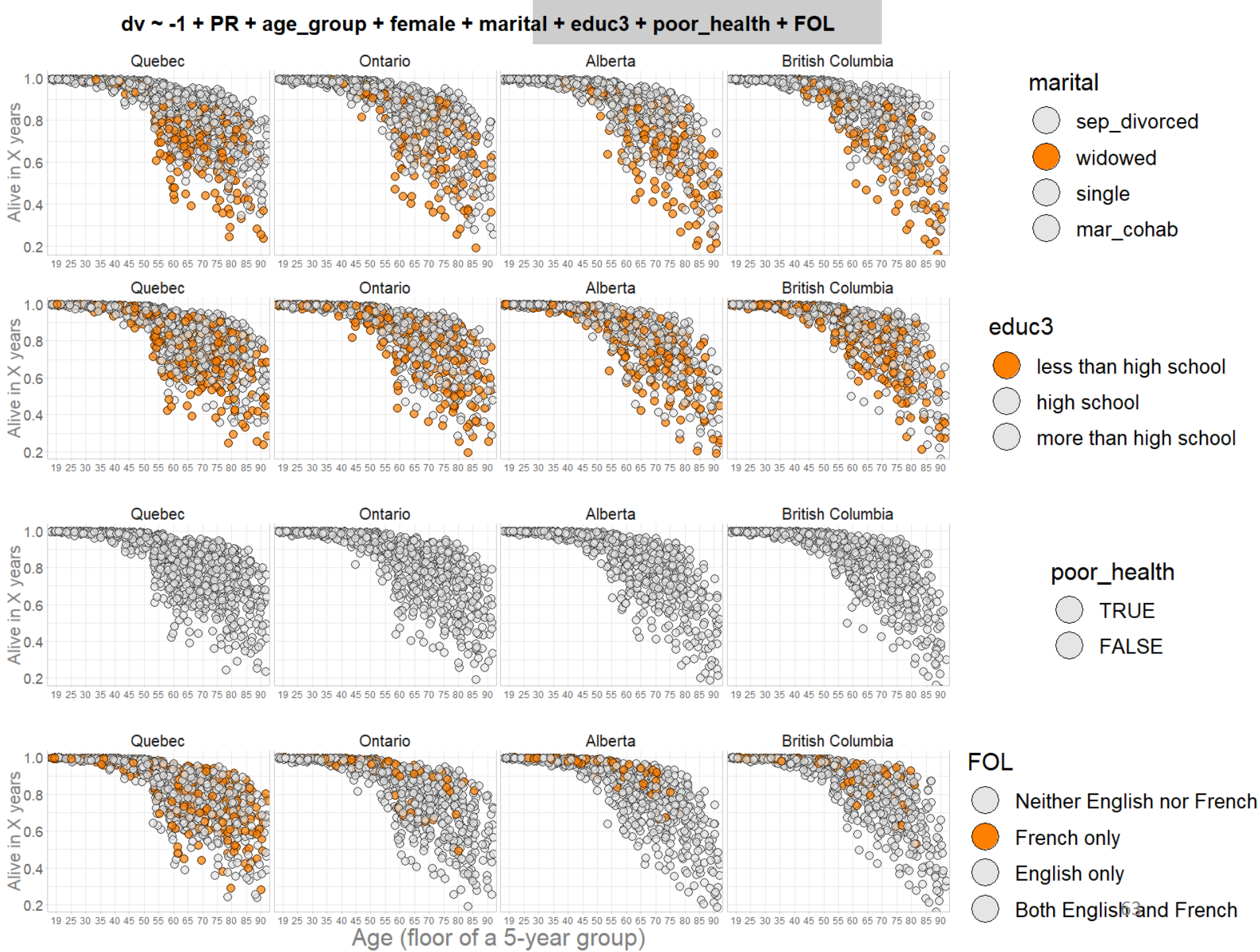
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province

Informed expectation

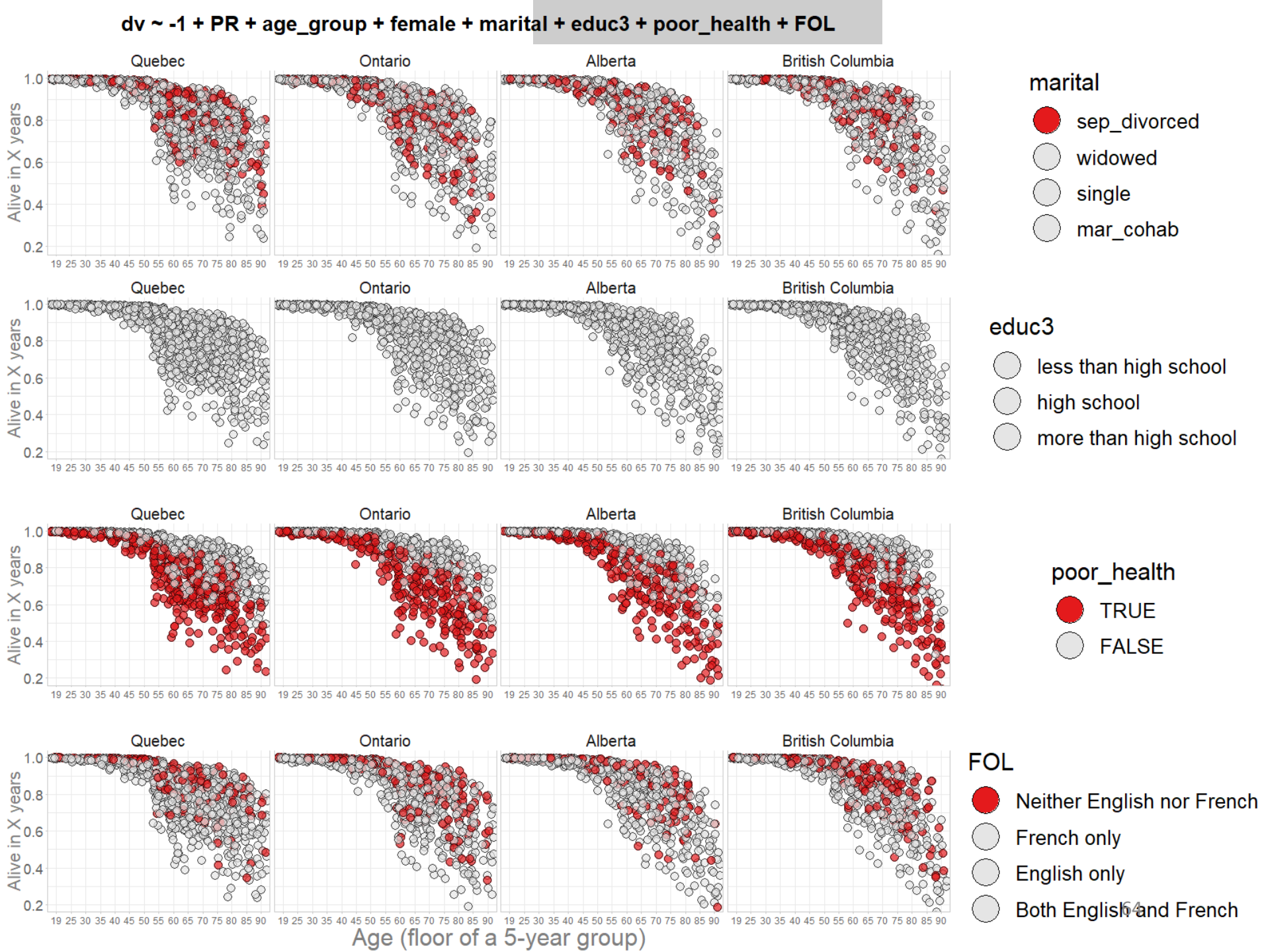
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



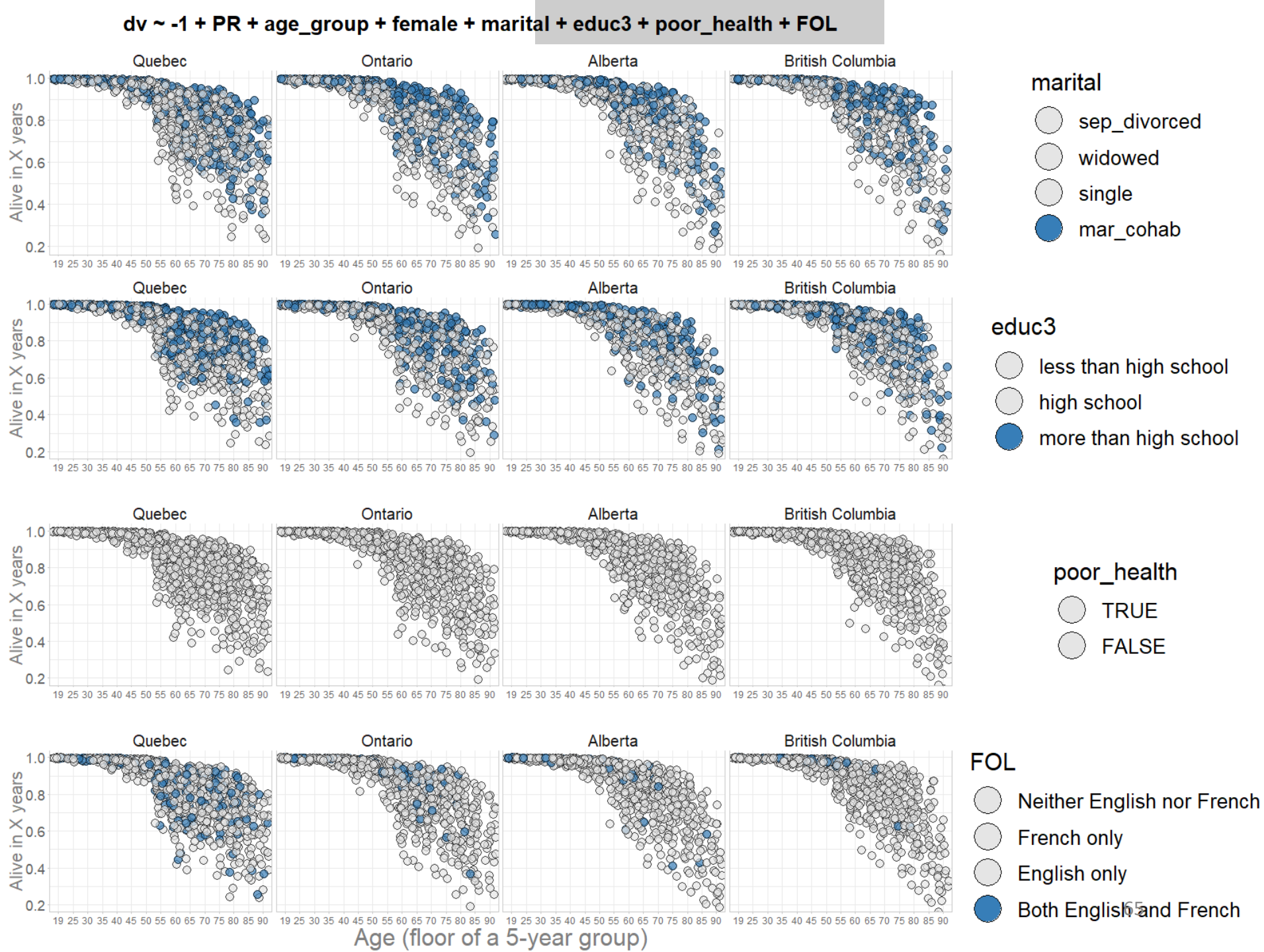
10,000 of 1st gen immigrants from each province

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



10,000 of 1st gen immigrants from each province



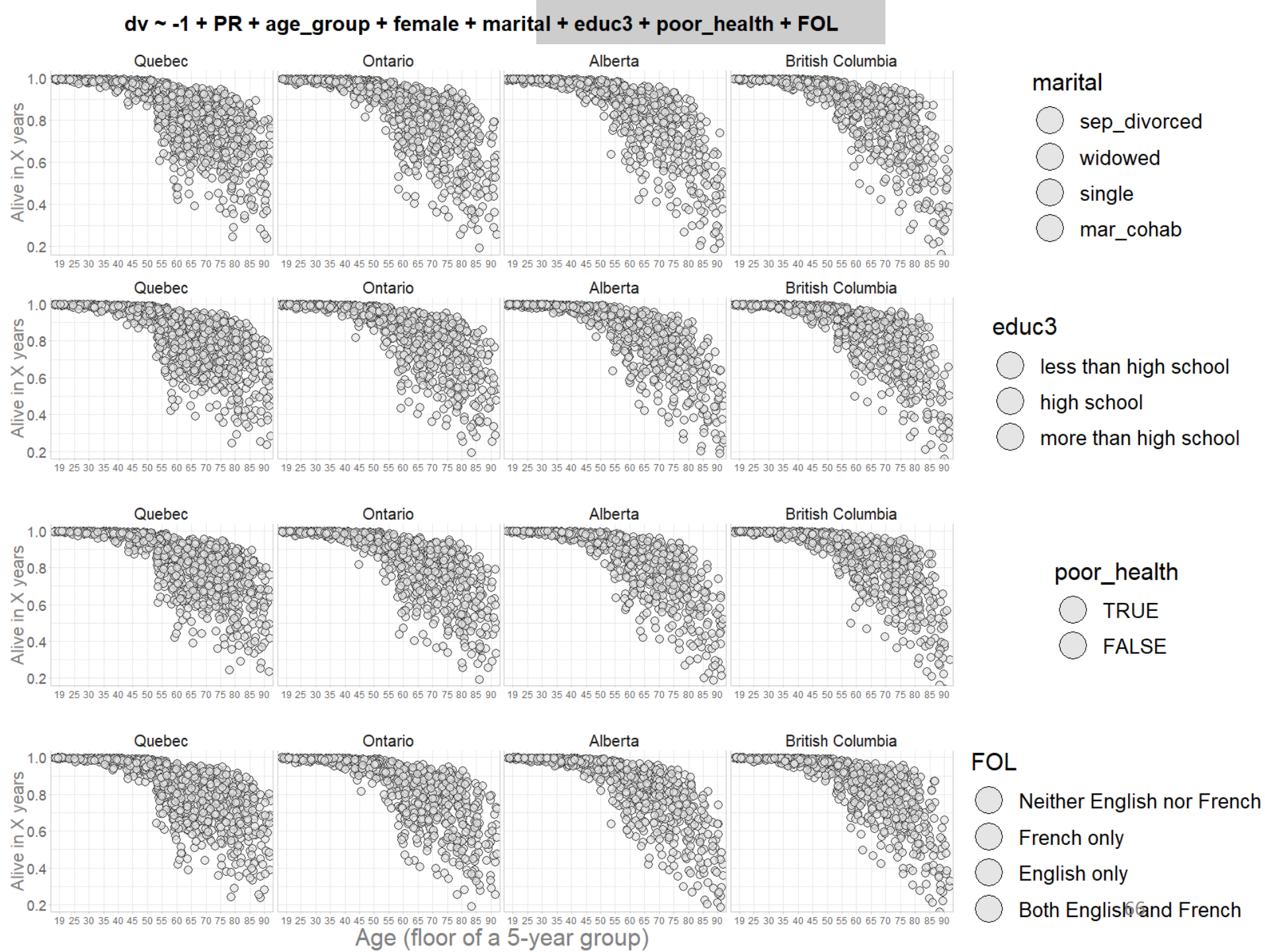
Informed expectation

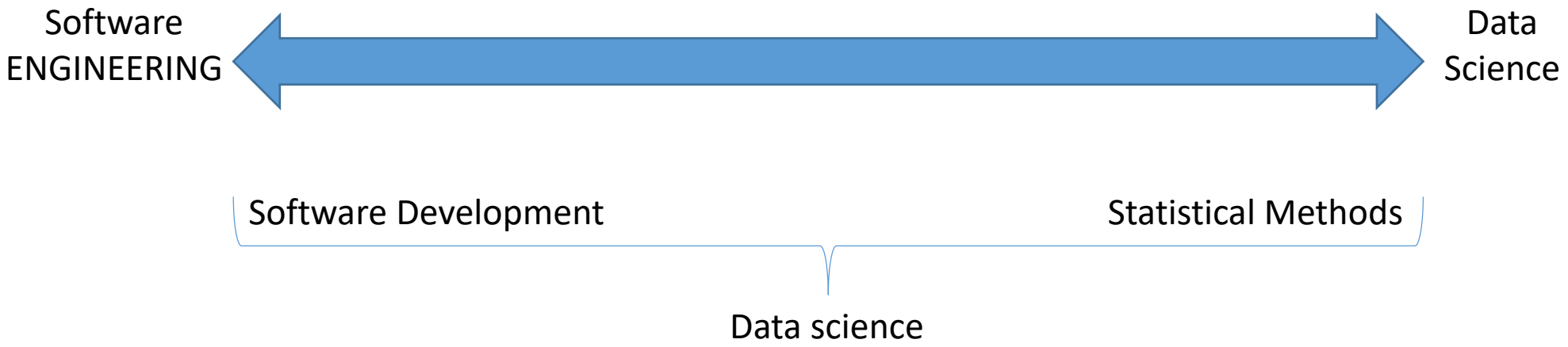
- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk

10,000 of 1st gen immigrants from each province

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk




Software
ENGINEERING
Data
Science
Software Development
Statistical Methods
Data science

Background – micro data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	ABDERR_synth	ABIDENT_synth	ADIFCLTY_synth	CITSM_synth	COWD_synth	DISABFL_synth	DISABIL_synth	DVISMIN_synth	FOL_synth	FPTIM_synth	GENSTPOB_synth	HCDD_synth	IMMDER_synth	LOINCA_synth	LOINCB_synth	MARST_synth	N
2	2	6	1	2	4	1	9	14	1	1	1	9	1	1	1	2	
3	2	6	1	2	4	1	9	14	1	1	1	8	1	1	1	2	
4	2	6	1	1	7	4	14	14	2	3	3	1	3	1	1	2	
5	2	6	1	1	4	1	9	14	1	2	3	2	3	1	2	4	
6	2	6	1	1	4	1	9	14	1	1	3	5	3	1	1	2	
7	2	6	1	1	7	1	9	14	2	3	3	1	3	1	1	2	
8	2	6	1	1	4	1	9	14	1	1	3	3	3	1	1	4	
9	2	6	1	1	4	1	9	14	1	1	3	9	3	1	1	2	
10	2	6	1	1	7	4	10	14	2	3	3	1	3	1	1	2	
11	2	6	1	1	4	1	9	14	1	1	3	6	3	1	1	4	
12	2	6	1	1	4	1	9	14	1	1	3	2	3	1	1	2	
13	2	6	1	1	4	1	9	14	3	1	1	2	1	1	2	4	
14	2	6	1	1	4	1	9	14	1	1	3	8	3	1	1	2	
15	2	6	1	1	4	1	9	14	1	1	3	9	3	1	1	4	
16	2	6	1	1	4	1	9	14	1	1	3	9	3	1	1	2	
17	2	6	1	1	4	1	9	2	2	1	3	3	3	1	1	4	
18	2	6	1	1	4	1	9	1	1	1	1	2	1	1	1	2	
19	2	6	1	1	4	1	9	2	1	1	1	7	1	1	1	2	
20	2	6	1	1	7	1	9	14	1	3	3	9	3	3	3	1	
21	2	6	1	1	4	1	9	14	1	1	3	2	3	1	1	4	
22	2	6	1	1	7	1	9	14	1	3	2	4	3	1	1	5	
23	2	6	1	1	3	1	9	14	1	1	3	2	3	1	1	2	
24	2	6	1	1	4	1	9	14	1	1	3	2	3	1	1	2	
25	2	6	1	1	4	1	9	14	1	1	3	8	3	1	1	2	
26	2	6	1	1	4	1	9	14	1	1	3	6	3	1	1	3	
27	2	6	1	1	4	1	9	14	1	1	3	2	3	1	1	2	
28	2	6	2	1	7	2	17	14	1	3	1	3	1	1	1	2	

github.com/andkov/ipdln-2018-hackathon

```
selected_provinces <- c("Alberta","British Columbia", "Ontario", "Quebec")
sample_size = 10000

# middle aged immigrants in british columbia
ds1 <- ds0 %>%
  dplyr::filter(PR %in% selected_provinces) %>%
  dplyr::filter(IMMDER == "Immigrants") %>%
  dplyr::filter(GENSTPOB == "1st generation - Respondent born outside Canada")
```

```
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                      DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
                      )
```

```
,educ3 = car::recode(
  HCDD, "
 'None'                                                                                 = 'less than high school'
;'High school graduation certificate or equivalency certificate'                       = 'high school'
;'Other trades certificate or diploma'                                                 = 'more than high school'
;'Registered apprenticeship certificate'                                               = 'more than high school'
;'College, CEGEP or other non-university certificate or diploma from a program of 3 months to less than 1 year'  = 'more than high school'
;'College, CEGEP or other non-university certificate or diploma from a program of 1 year to 2 years'             = 'more than high school'
;'College, CEGEP or other non-university certificate or diploma from a program of more than 2 years'            = 'more than high school'
;'University certificate or diploma below bachelor level'                             = 'more than high school'
;'Bachelors degree'                                                                    = 'more than high school'
;'University certificate or diploma above bachelor level'                             = 'more than high school'
;'Degree in medicine, dentistry, veterinary medicine or optometry'                     = 'more than high school'
;'Masters degree'                                                                      = 'more than high school'
;'Earned doctorate degree'                                                             = 'more than high school'
")
```

```
,marital = car::recode(
  MARST, "
   'Divorced'                              = 'sep_divorced'
  ;'Legally married (and not separated)'   = 'mar_cohab'
  ;'Separated, but still legally married'  = 'sep_divorced'
  ;'Never legally married (single)'        = 'single'
  ;'Widowed'                               = 'widowed'
")
```

```
# All colors are in
increased_risk_2  <- "#e41a1c"  # red      - further increased risk factor
increased_risk_1  <- "#ff7f00"  # organge  - increased risk factor
reference_color   <- "#4daf4a"  # green    - REFERENCE  category
descreased_risk_1 <-"#377eb8"   # blue     -  descreased risk factor
descreased_risk_2 <- "#984ea3"  # purple   - further descrease in risk factor
```