Applied Bayesian Modeling module 9:
Expanding our model universe

Leontine Alkema, lalkema@umass.edu
Fall 2022

Introduction

- So far, we discussed model specifications based on a normal likelihood function, also referred to as data model: $y_i|\mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$, or, equivalently, $y_i = \mu_i + \varepsilon_i$ with $\varepsilon_i|\sigma \sim N(0, \sigma^2)$.
- We discussed specifications of μ_i all the way to hierarchical regression models, i.e. from module 9

$$\mu_i \quad = \quad \alpha + \eta_{0,j[i]} + (\beta_1 + \eta_{1,j[i]})x_i + \beta_2 u_{j[i]} + \beta_3 u_{j[i]} x_i.$$

  The specification of the unknown parameter of interest is also referred to as the process model.
- But what if...

What if ...

- What if data are not normal w. constant variance but we observe
  - binary outcomes or counts
  - more complex data generating mechanisms, e.g., data subject to unknown reporting errors that vary with characteristics of the data collection process
- What if outcome of interest μ_i is NOT a linear function of covariates but
  - depends on covariates in a non-linear way
  - depends on its own past (when monitoring/forecasting an outcome over a time period)
  - is best described with a set of equations and variables that depend on one another (infectious disease modeling, abortion accounting model)
- The good news:
  - Lots of flexibility to fit a variety of Bayesian models using software such as Stan
  - Going Bayesian allows for specification of variety of data and process model structures and it produces uncertainty assessments for parameters or unobserved units that account for uncertainty associated with data and additional parameters

Outline of modules

- This module: Introduce some alternative models, using real data examples
  - Examples: generalized linear hierarchical models (logistic regression); estimating abortion incidence from 1990 to 2019 for all countries in the world
  - Get inspired about what you could do for your project
- Next modules:
  - How to fit models using Stan, choice of priors when specifying your own models
  - Bayesian workflow; Model checking and validation

Example 1: well switching in Bangladesh (GH 5.4-5.6)

- Wells in Bangladesh are contaminated with natural arsenic.
- In a study, wells were labeled as safe or unsafe.
- Researchers returned and recorded which households using unsafe wells had switched wells, as well as distance to nearest safe well (and other things).
- Outcome of interest: probability of switching π_i for household i.
- Data $y_i = 0$, if household i switched, 0 otherwise.
- Predictor variable d_i= distance to the nearest safe well.

Logistic regression: main idea

- Outcome of interest: probability of success π_i and how it relates to a set of covariates, where examples of "success" are
  - whether household switches wells
  - coin lands heads, person does not get the flu this season, ...
- Data y_i are binary outcomes (success 1 or failure 0), or the number of successes out of a fixed number of independent trials.
- Basic set-up for data model

$$\begin{aligned} y_i|\pi_i &\sim Bern(\pi_i), \text{ OR} \\ y_i|\pi_i &\sim Bin(n_i, \pi_i), \text{ if total number of trials is } n_i. \end{aligned}$$

- Parameter π_i needs to be constrained to be in (0,1), can use a transformation such as logit:

$$\begin{aligned} logit(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \text{ some function of covariates, e.g.} \\ logit(\pi_i) &= \beta_0 + \beta_1 x_i. \end{aligned}$$

Example 1: continued

- Data: $y_i = 0$, if household i switched, 0 otherwise.
- Predictor variable d_i= distance to the nearest well.
- Basic set-up:

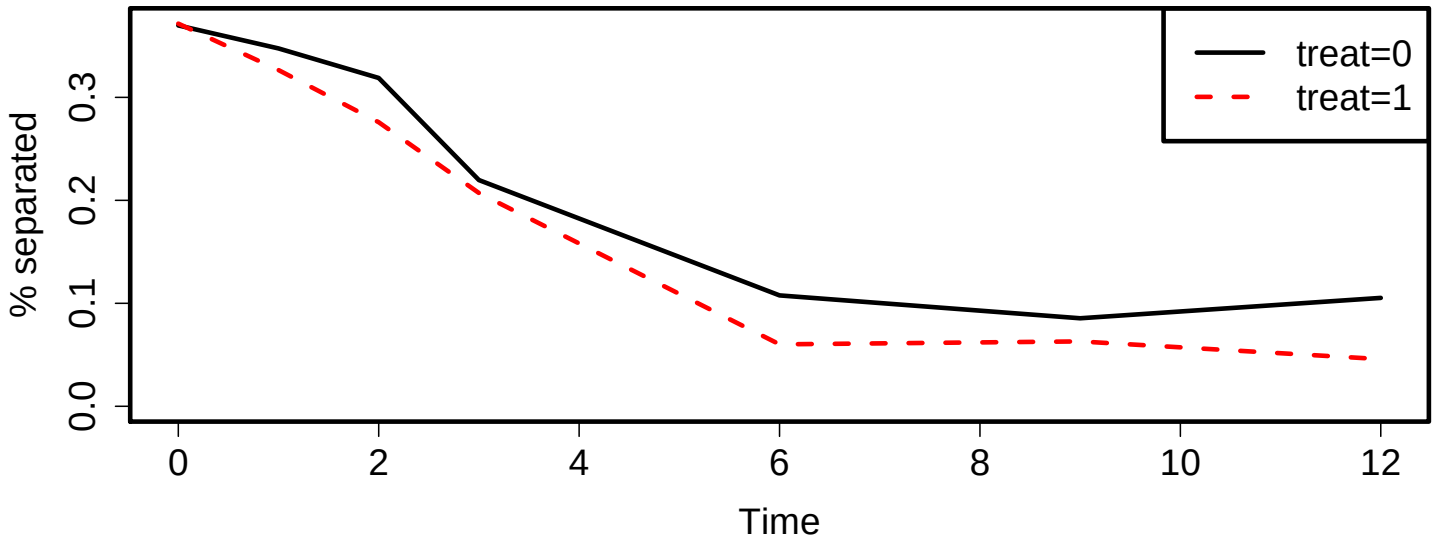
$$\begin{aligned} y_i|\pi_i &\sim Bern(\pi_i), \\ logit(\pi_i) &= \beta_0 + \beta_1(d_i - \bar{d}). \end{aligned}$$

- We can fit this model in `brm`

Logistic regression: a more advanced example

- Toenail RCT, taken from LL (Lesaffre and Lawson, 2012) Ch.9, with reference De Backer et al (1996).
- Set-up:
  - A randomized controlled clinical trial comparing two oral treatments for toenail dermatophyte onychomycosis (infection).
  - One of two oral medications was used: Itraconazol 250 mg daily (treat=0) or Lamisil 250 mg daily (treat=1).
  - The patients received treatment for twelve weeks and were evaluated at 0, 1, 2, 3, 6, 9 and 12 months.
- Here the response of interest is the binarized degree of onycholysis (nail separation) of a subgroup of 294 patients, with
  - 0 = no/mild separation, 1 = moderate or severe separation.

Toenail RCT



- % separated refers to proportion of individuals with moderate or severe separation
- Q: how to estimate the (difference in) treatment effect?
- Next: simple-ish linear set-up

Toward a candidate model set-up

- Data and separation proportion:

$$\begin{aligned} y_i|\pi_i &\sim Bern(\pi_i), \\ logit(\pi_i) &= \text{ some function of time } t_i \text{ and treatment } x_i \text{ ,} \end{aligned}$$

  where for observation i:
  - $y_i = 0$ if no/mild separation, $1 =$ moderate or severe separation.
  - $t_i =$ measurement time,
  - $x_i =$ treatment variable, with $x_i = 1$ for treatment L, 0 otherwise.
- Simple linear set-up:
  assume for someone NOT on treatment that

$$logit(\pi_i) = \alpha + \beta t_i,$$

  and assume a constant difference in logit(probs) between treated and untreated:

$$logit(\pi_i) = \alpha + (\beta + \gamma) t_i,$$

  where γ represents the treatment effect.

Toward a candidate model set-up (ctd)

- The combined model can be written as:

$$\begin{aligned} y_i|\pi_i &\sim Bern(\pi_i), \\ logit(\pi_i) &= \alpha + \beta t_i + \gamma t_i x_i. \end{aligned}$$

(x_i is not added as a first order term because treatment has not yet started at $t = 0$.)

- What does this model ignore?
  That observations are grouped by person!
- Let $j[i]$ refer to the patient index of observation i (so one group = one patient). We can allow for person-specific curves as follows:

$$\begin{aligned} y_i|\pi_i &\sim Bern(\pi_i), \\ logit(\pi_i) &= \alpha_{j[i]} + \beta_{j[i]} t_i + \gamma t_i x_i. \end{aligned}$$

Estimating person-specific intercepts and slopes

- As for the radon data, we can use a bivariate normal distribution for the person-specific intercepts and slopes:

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} | \mu_\sigma, \mu_\beta, \sigma_\alpha, \sigma\beta, \rho \sim N_2 \left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right).$$

Model fitting

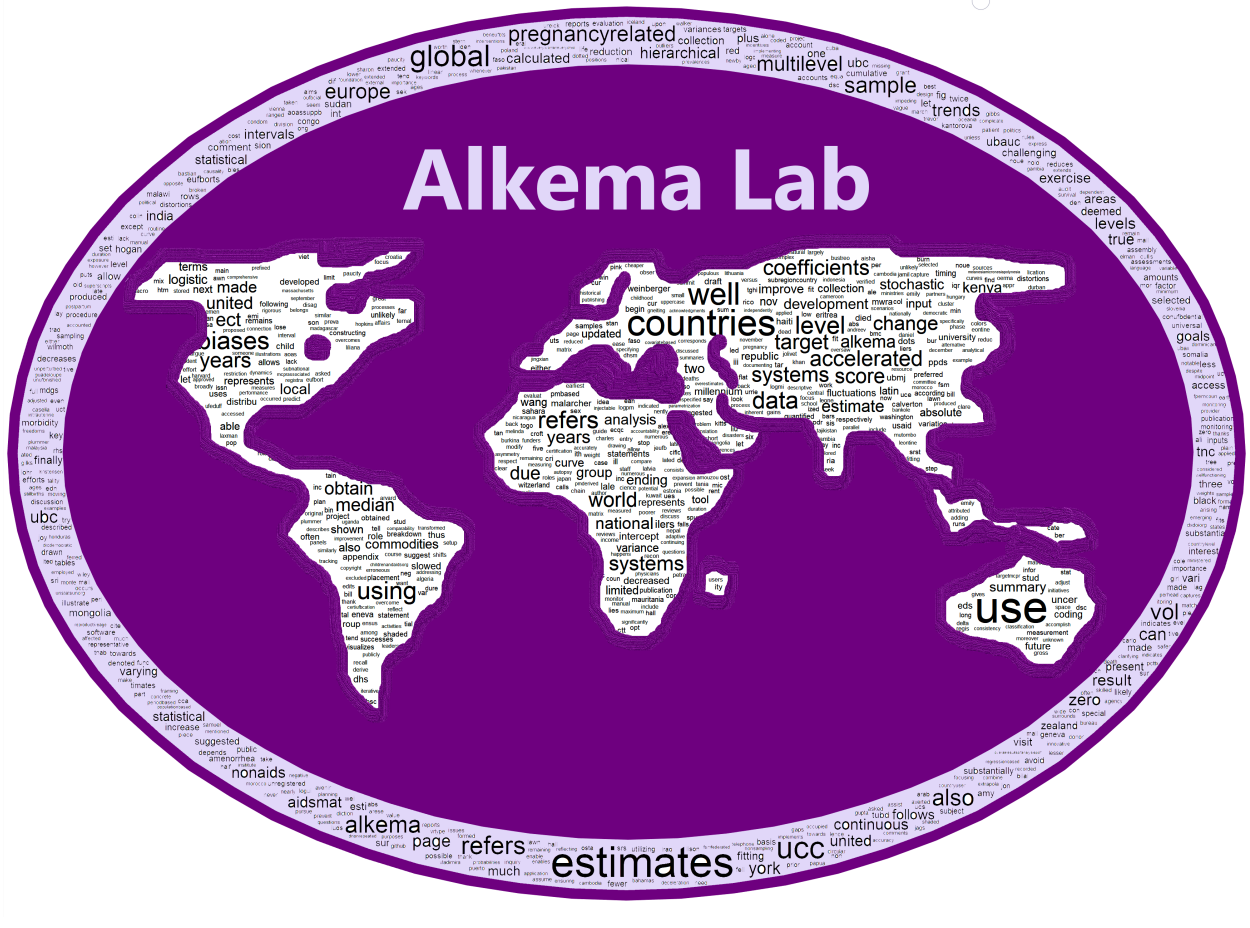
- Main finding when

$$\begin{aligned} y_i|\pi_i &\sim Bern(\pi_i), \\ logit(\pi_i) &= \alpha_{j[i]} + \beta_{j[i]}(t_i - \bar{t}) + \gamma t_i x_i, \end{aligned}$$

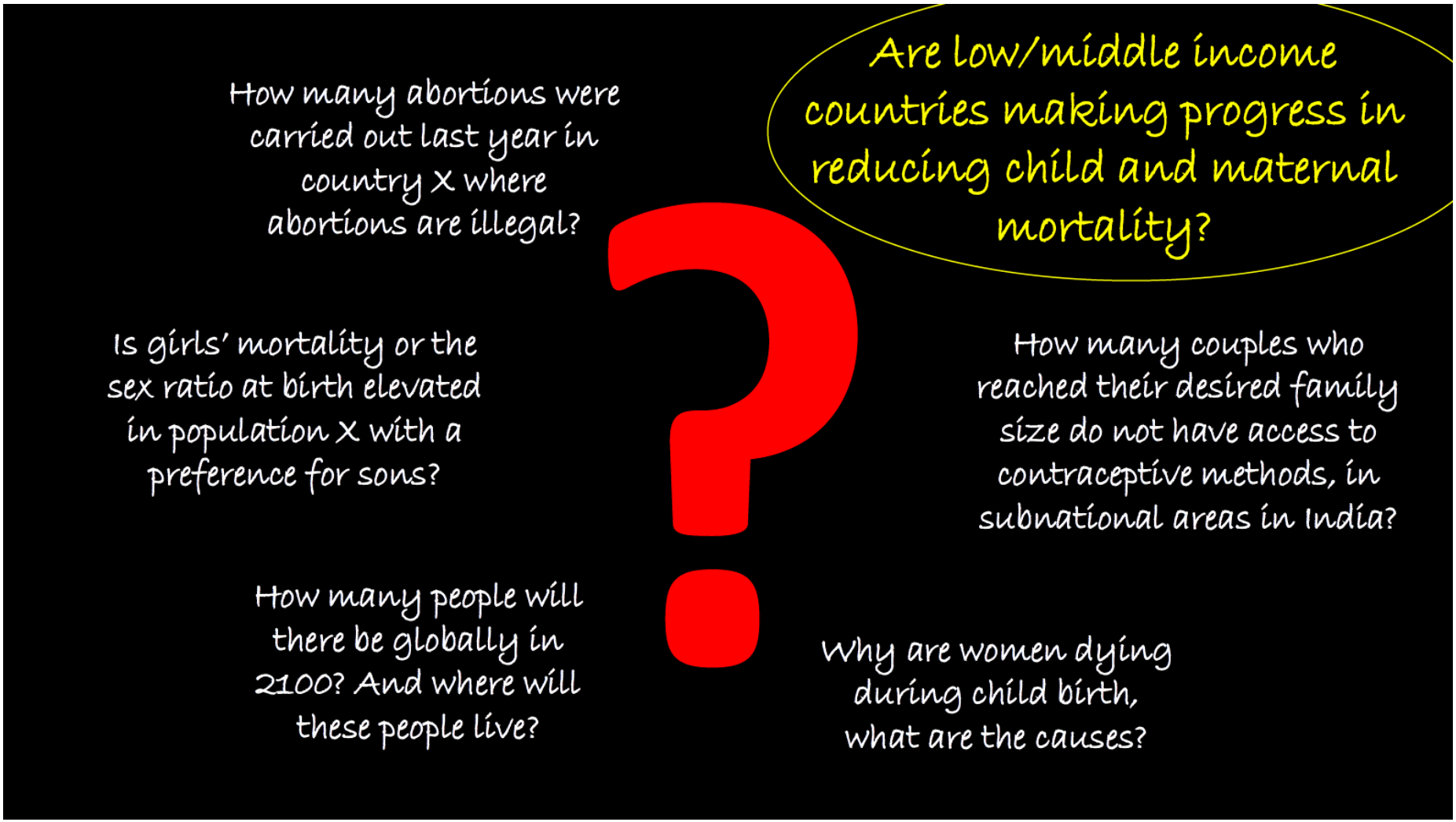
  γ is estimated at -0.40 with 95% CI (-0.51, -0.04), hence there is a significant treatment effect.
- Interpretation:
  - For a person with $x_i = 0$ (no treatment), the relative change in odds from month t to $t+1$ is $\exp(\beta_{j[i]})$
  - If that person would have received the treatment, so if $x_i = 1$, the estimated relative change in odds from t to $t+1$ is $\exp(\beta_{j[i]} + \gamma) = \exp(\beta_{j[i]})\exp(\gamma)$, with $\exp(\gamma) = 0.67$.
  - Hence the treatment is associated with a 33% faster decrease in odds of nail separation for each month.

Bayesian modeling of population health indicators in Alkema lab @ UMass Amherst

- See `https://leontinealkema.github.io/alkema_lab/about.html`

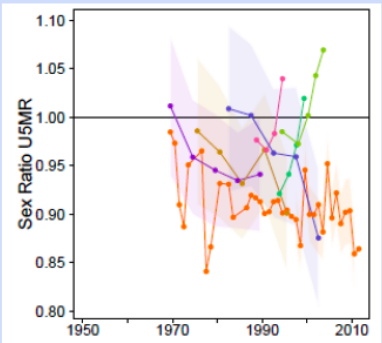


The questions we try to answer...

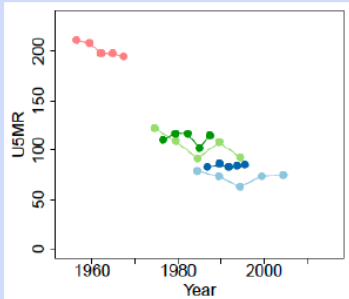


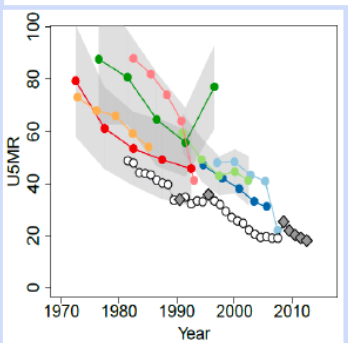
Why we need models to answer these questions…

Sex ratio of child mortality in India

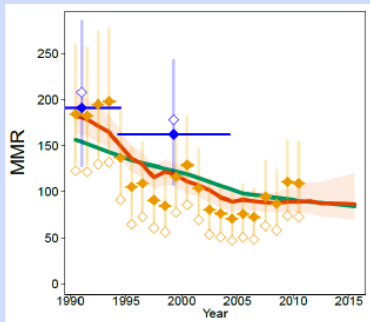


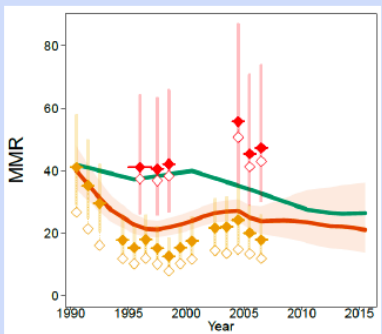
Child mortality in Papua New Guinea and Kazakhstan

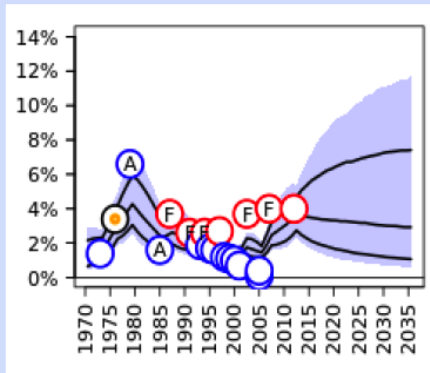




Maternal mortality in Ecuador and Thailand



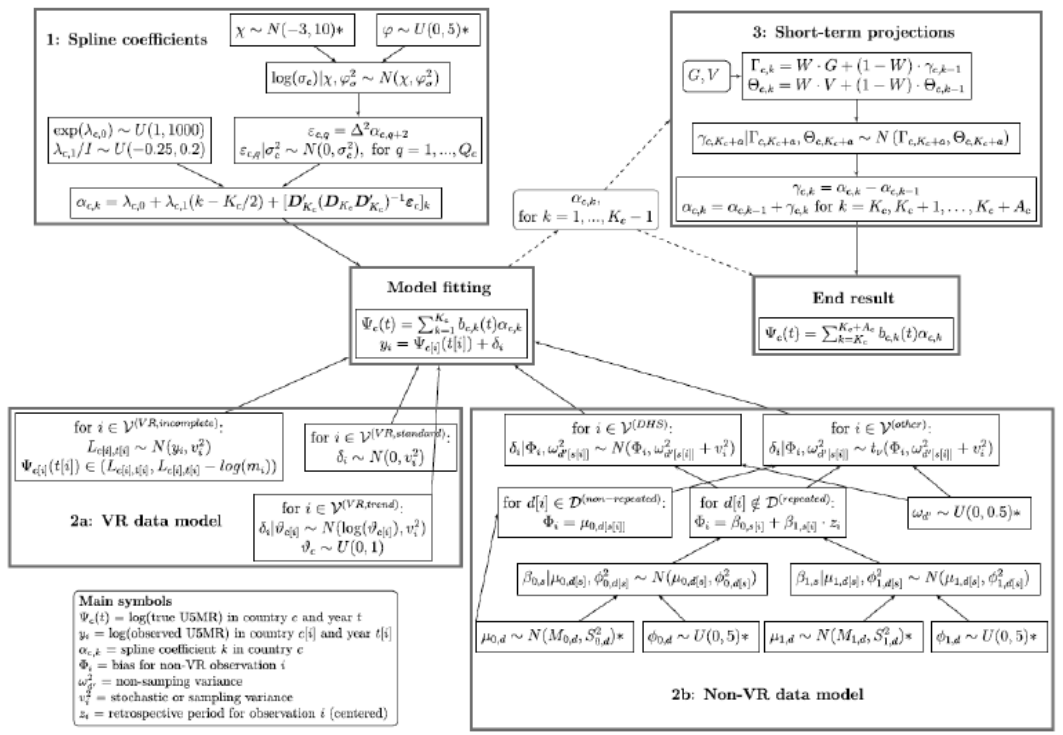




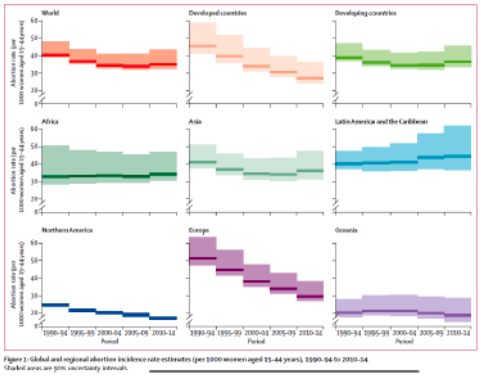
Use of traditional contraceptive methods in Indonesia

Bayesian model development

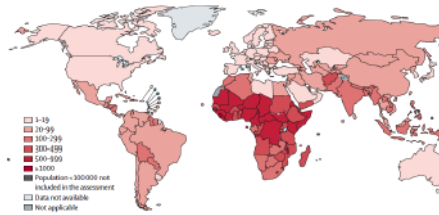
- using a variety of techniques (hierarchical models, penalized splines regression models, time series (ARIMA) modeling, Bayesian melding),
- with an emphasis on appropriate data models to include wide variety of data sources.



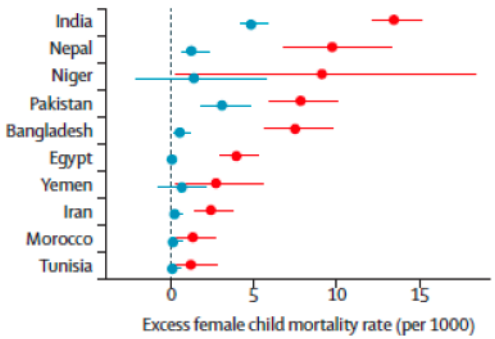
How many abortions were carried out last year in country X where abortions are illegal?



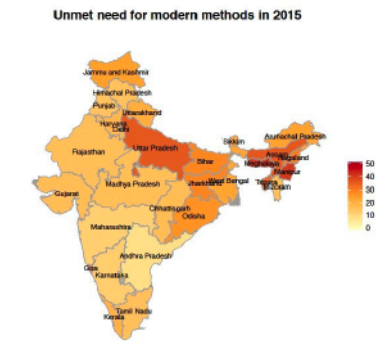
Are developing countries making progress in reducing child and maternal mortality?



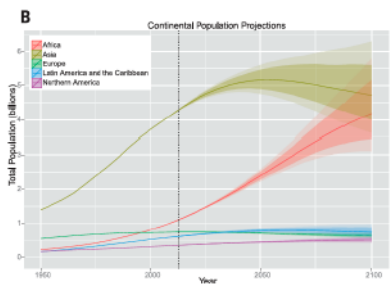
Is girls' mortality or the sex ratio at birth elevated in population X with a preference for sons?



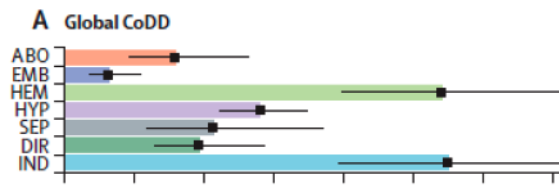
How many couples who reached their desired family size do not have access to contraceptive methods, in subnational areas in India?



How many people will there be globally in 2100? And where will these people live?



Why are women dying during child birth, what are the causes?



Example: Estimating incidence of abortion and unintended pregnancies

- Goal: estimate abortion incidence for all countries, from 1990 to 2019
- Problem: data are limited and may be subject to substantial errors

Illustration of data for a high-income country (HIC)

Data in colored symbols; abortion rate = # abortions per 1,000 women

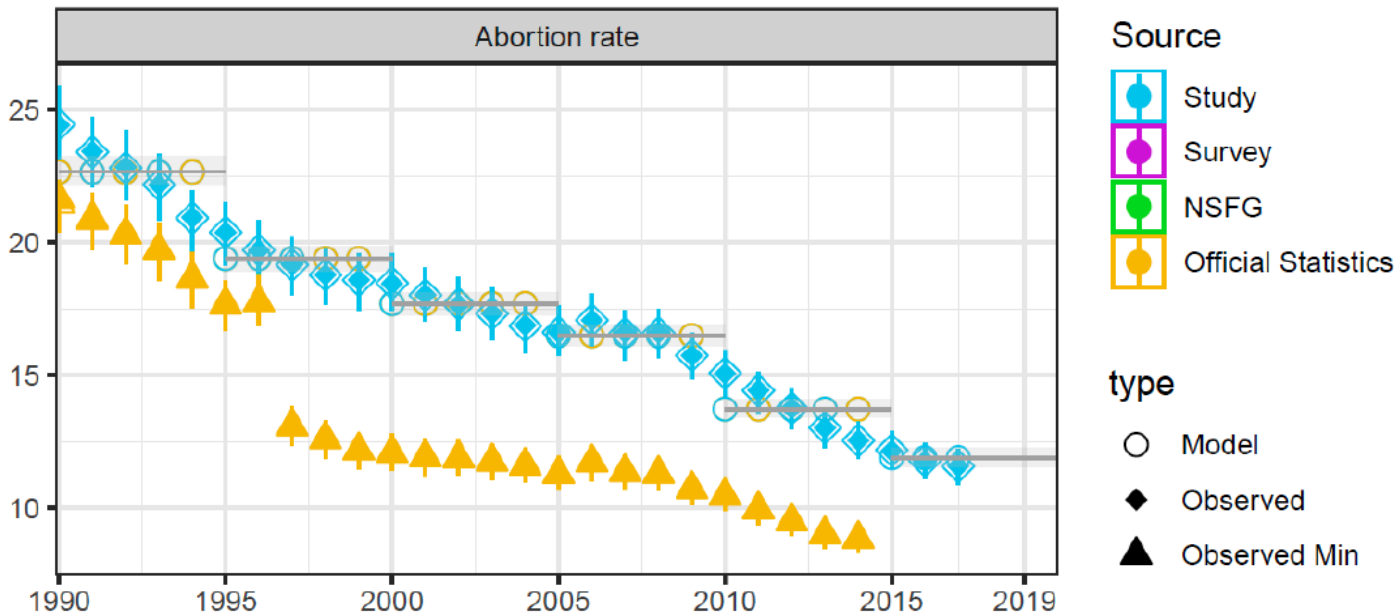
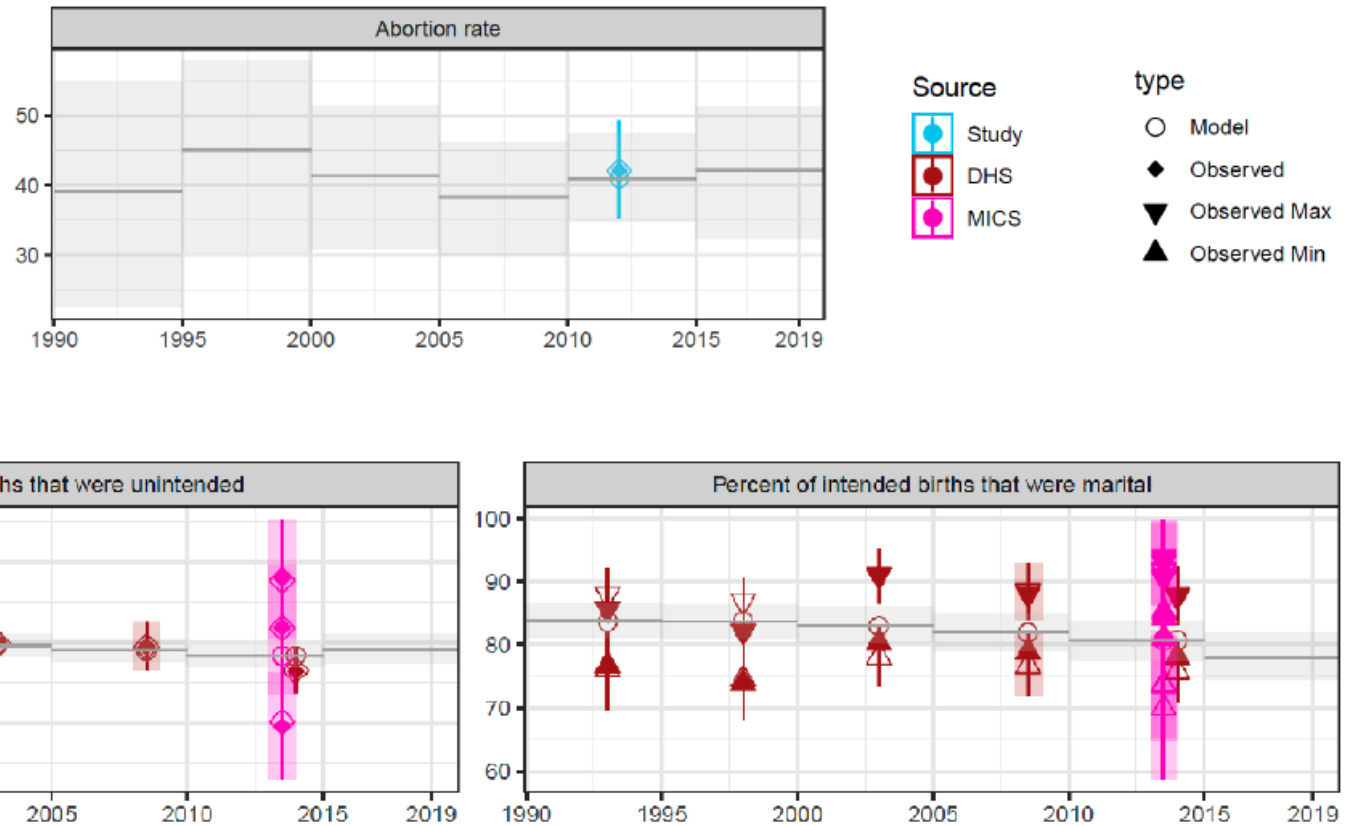


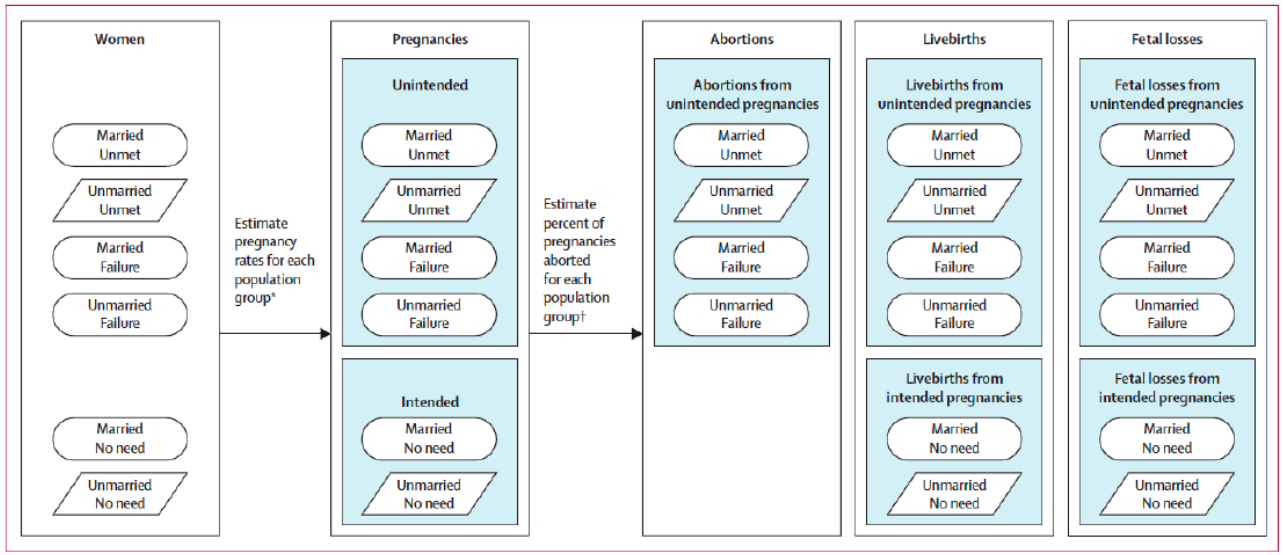
Illustration of data for a middle-income country (MIC)

Data in colored symbols; abortion rate = # abortions per 1,000 women



Estimating incidence of abortion and unintended pregnancies

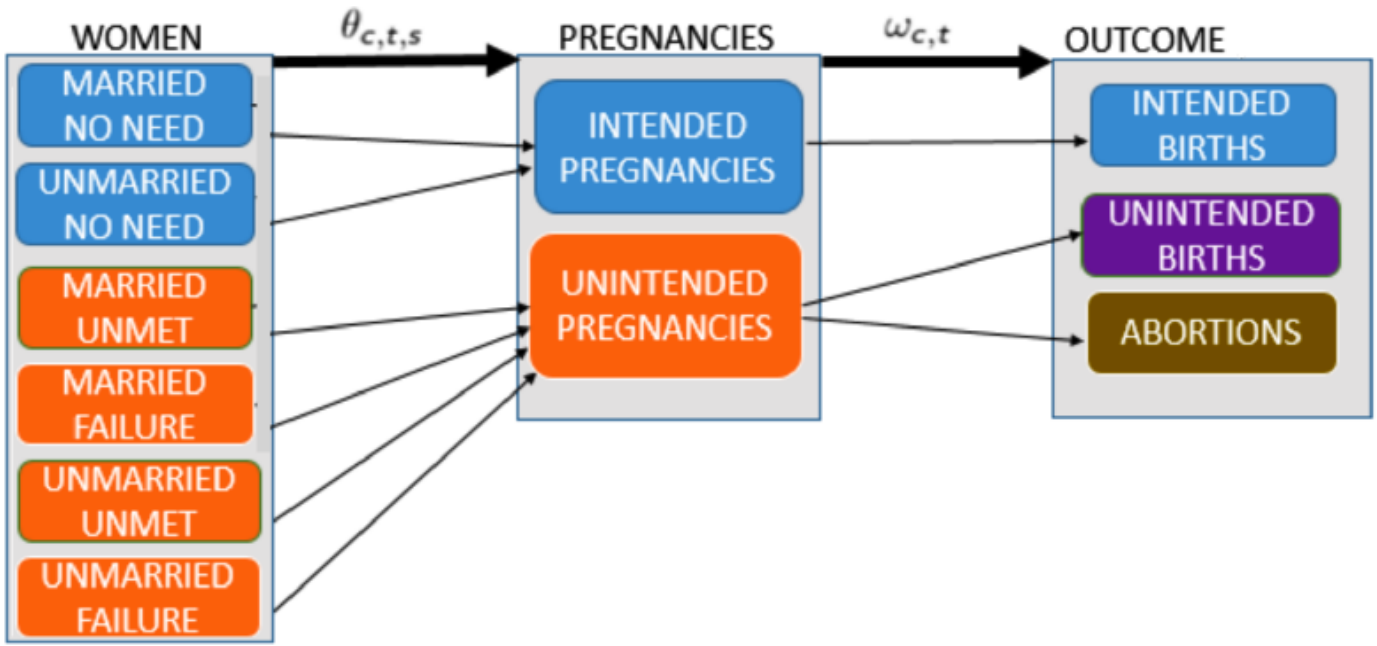
- Goal: estimate abortion incidence for all countries, from 1990 to 2019
- Problem: data are limited and may be subject to substantial errors
- Approach: Bayesian accounting model to estimate unintended pregnancies and abortions (Bearak et al., 2020a)



Bearak, Popinchalk, Ganatra, Moller, Tuncalp, Beavin, Kwok, Alkema (2020a).
Unintended Pregnancy and Abortion by Income, Region, and the Legal Status of Abortion: Estimates from a Comprehensive Model for 1990-2019. The Lancet Global Health 8 (9): e1152-61.

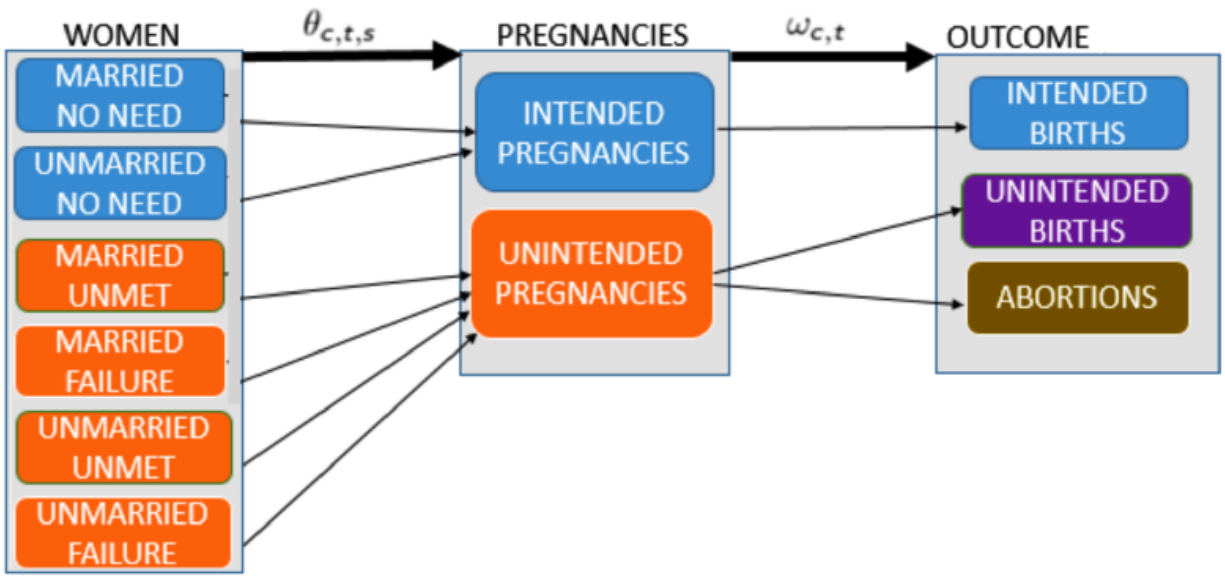
Simplified summary of the Bayesian accounting model

- For each country-period we set up a demographic accounting model:
  - Left: Subgroups of women based on (simplified)
    - marital status; contraceptive use yes/no; wants a child yes/no
  - Middle: Pregnancies by intention status
  - Right: Births by intention/union status; abortions.
- Parameters:
  - θ_s = subgroup-specific pregnancy rates
  - ω = propensity to abort unintended pregnancies



Simplified summary of the Bayesian accounting model (ctd)

- Model fitting for all country-periods:
  - For all country-periods c, t: estimates of subgroups and total births
  - For a **subset** of country-periods c, t only:
    Data on births by intention/union status, abortions
  - Specify data models to account for reporting issues



Bayesian multilevel time series models

- To produce estimates for all country-periods:
  - Bayesian multilevel time series models for the pregnancy rates $\theta_{c,t,s}$ and propensity to abort $\omega_{c,t}$ (Bearak et al., 2020b)
- Example multilevel time series model for $\theta_{c,t,s}$:
  random walk with region-specific drift $\Delta_{r[c],t,s}$:

$$\theta_{c,t,s}|\theta_{c,t-1,s}, \Delta_{r[c],t,s} \sim N\left(\theta_{c,t-1,s} + \Delta_{r[c],t,s}, \sigma^2\right),$$

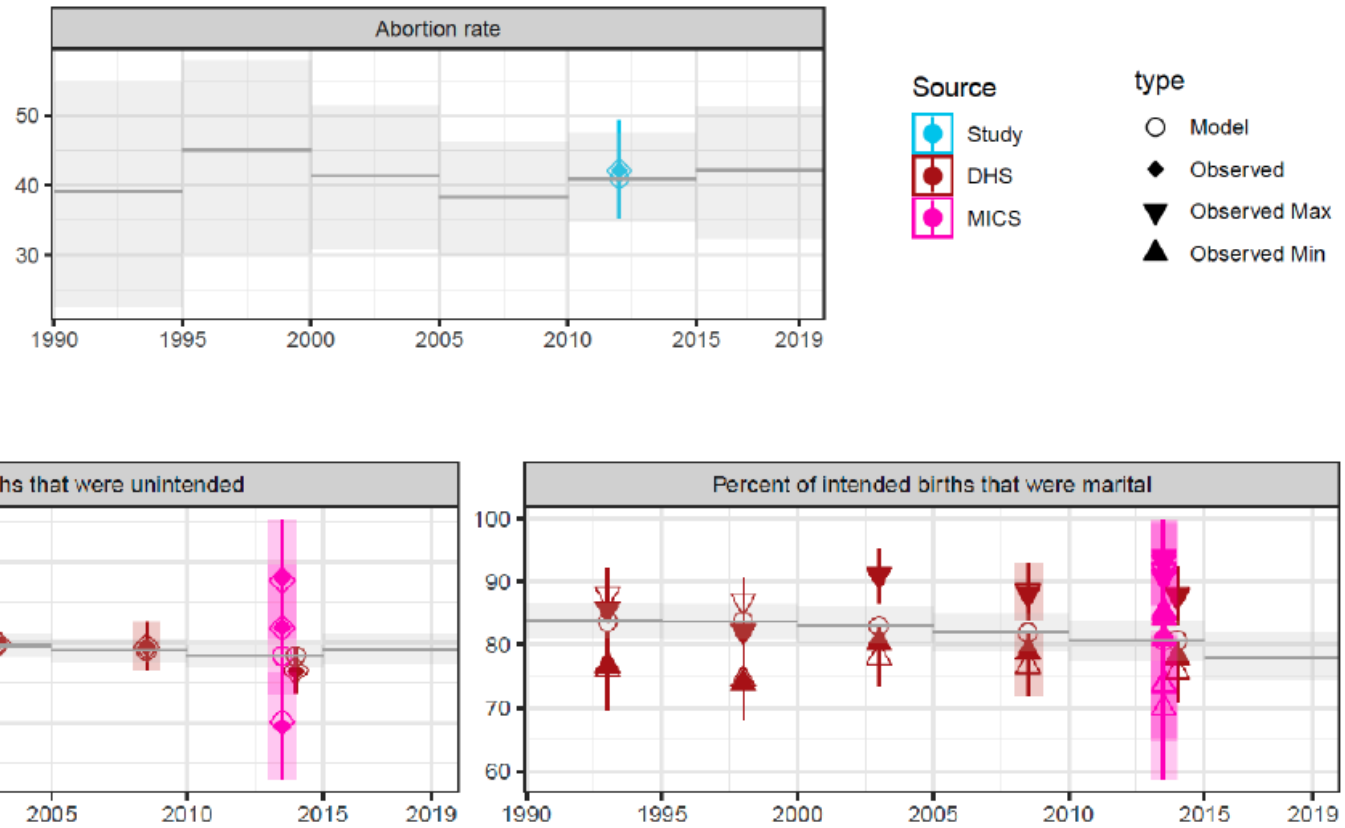
  where $r[c]$ = region of country c.
- Then
  - Data on period t^* can inform other periods $t \neq t^*$;
  - Data on country c^* can inform other countries c in a region

  $\Rightarrow$ Produce estimates for all country-periods, including those without intention or abortion data

Bearak, Popinchalk, Ganatra, Moller, Tuncalp, Beavin, Kwok, Alkema (2020b).
Global estimation of unintended pregnancy and abortion using a Bayesian hierarchical random walk model. See https://arxiv.org/abs/2007.09246.

Illustration of data for a middle-income country (MIC)

Data in colored symbols; abortion rate = # abortions per 1,000 women



Extending our model universe: summary

- The good news:
  - Lots of flexibility to fit a variety of Bayesian models using software such as Stan
  - Going Bayesian allows for specification of variety of data and process model structures and it produces uncertainty assessments for parameters or unobserved units that account for uncertainty associated with data and additional parameters
- You get a chance to develop a Bayesian model for your outcomes of interest in the project
- Next modules:
  - How to fit models using Stan, choice of priors when specifying your own models
  - Bayesian workflow; Model checking and validation