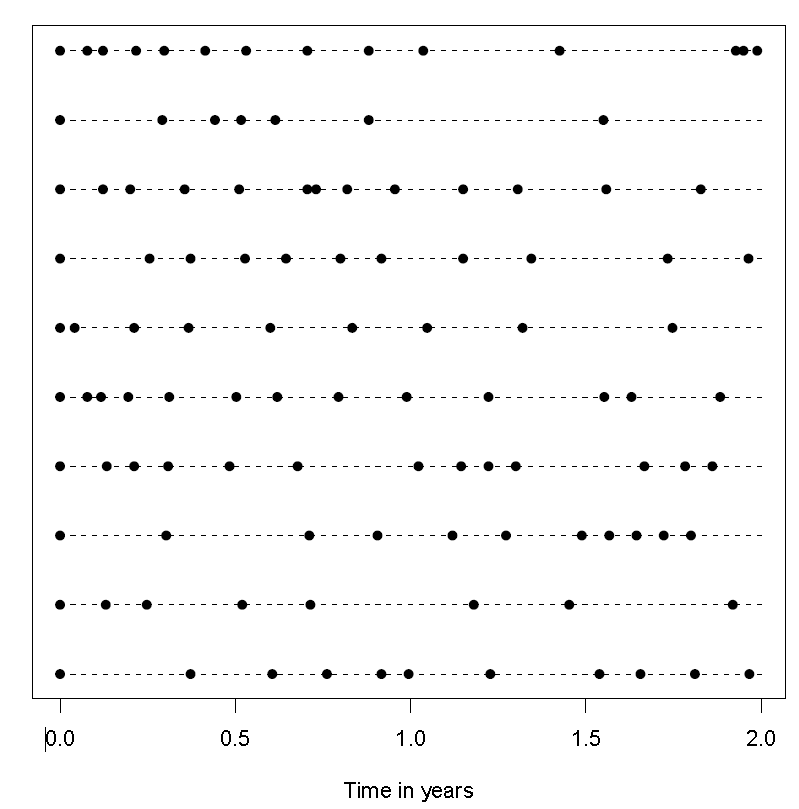
Analysing longitudinal data when the visit times are informative

Eleanor Pullenayegum, PhD
Scientist, Hospital for Sick Children
Associate Professor, University of Toronto
eleanor.pullenayegum@sickkids.ca

The Problem



- Trials
- Prospective cohorts
- Retrospective cohorts

Learning objectives

- Explain why irregular observation can be a problem
- Understand different visiting patterns
- Choose an appropriate analysis for a given visiting pattern

Informative visit times

- Why does this happen?
- When is it a problem?
- How can we handle it?
  - Standard methods
  - Inverse-intensity weighting
  - Semi-parametric joint models
- Study design

Why?

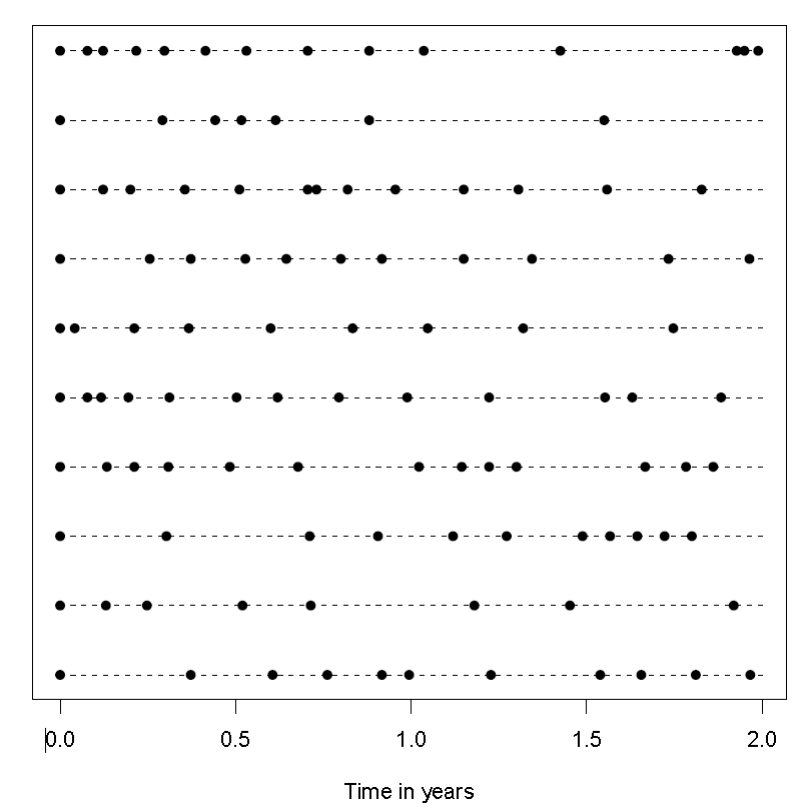
- Repeated measures studies with variation in visit times
- Studies without “study visits”
  - Chart reviews, etc.

Example – Juvenile Dermatomyositis (JDM)

- Rare disease affecting skin & muscles
- Followed children (n=92) in a clinic at Sick Kids
- Follow-up was as part of clinical care
  - Standard of care: 0, 2, 4, 6, 10, 14, 22, 30, and 42 weeks, and thereafter every 2 to 3 months
- More frequent follow up after disease flares
- Aim: describe disease activity over time

When is it a problem?

Number of visits varies from 3 to 14, mean of 10



Problematic if visit times are related to outcome

Notation

- Outcome $Y_i(t)$
- Visits at $T_{i1}, T_{i2}, \ldots$
- Counting process for visits $N_i(t)$ with intensity $\lambda_i(t)$
- Covariates $X_i(t)$, auxiliary covariates $Z_i(t)$
- Usually interested in regression models

$$E(Y_i(t) \mid X_i(t)) = g^{-1}(\beta_0(t) + X_i(t)\beta)$$

- Potentially problematic if N_i and Y_i are dependent.

Visit Schemes

Visit times can be…

- Independent of outcomes
  - Visiting completely at random (VCAR)
- Independent of outcomes given previously observed data
  - Visiting at random (VAR)
- Not independent of outcomes given previously observed data
  - Visiting not at random (VNAR)

Visit Processes Classification

- Visiting completely at random

$$E(\Delta N_i(t) \mid Y_i, X_i, Z_i) = E(\Delta N_i(t))$$

- Visiting at random

$$E(\Delta N_i(t) \mid Y_i, X_i, Z_i, N_i(s) : s < t)$$

$$= E(\Delta N_i(t) \mid Y_i^{obs}(s), X_i^{obs}(s), Z_i^{obs}(s), N_i(s) : s < t)$$

- Visiting not at random

$$E(\Delta N_i(t) \mid Y_i, X_i, Z_i, N_i(s) : s < t)$$

$$\neq E(\Delta N_i(t) \mid Y_i^{obs}(s), X_i^{obs}(s), Z_i^{obs}(s), N_i(s) : s < t)$$

Examples

Protocol

- Fixed visits (e.g. clinical trials) (VCAR)
- History-dependent protocol visits (VAR)
- Physician-driven visits (VAR)
- Patient-driven visits (VNAR)

Deviation

- Unrelated to outcomes
- Dependent on observed factors
- Dependent on unobserved factors

Analytical Solutions

- Standard analyses
- Inverse-weighting methods
- Semi-parametric joint models
- Parametric joint models

When to use what?

Table 1. Validity of analytic methods for various visit process models.

	Measured covariates for the visit process					Latent covariates for the visit process		Analytic model					
Visit process	No covariates permitted	Must also appear in outcome model	Past outcomes	No restrictions	May covariates be time dependent?	Correlated with outcome process?	May latent covariates be time varying?	Mixed models	GEE	Lin-Ying	Inverse-intensity weighted GEE	Semiparametric joint models	Parametric joint models
Regular	–	–	–	–	–	–	–	✓	✓	✓	✓	✓	✓
Visiting completely at random	•				N/A	No	Yes	✓	✓	✓	✓	✓	✓
Visiting at random		•			Yes	No	Yes	✓	✓	✓	✓	✓	✓
			•		Yes	No	Yes	✓	×	×	✓	×	✓
				•	No	No	Yes	×	×	×	✓	✓	✓
				•	Yes	No	Yes	×	×	×	✓	×	✓
Random-effect-dependent visits (special case of visiting not at random)		•			Yes	Yes	No	×	×	×	×	✓†	✓
				•	No	Yes	No	×	×	×	×	✓†	✓
				•	Yes	Yes	No	×	×	×	×	×	✓
				•	No	Yes	Yes	×	×	×	×	×	✓‡

Measured covariates that are predictive of visit frequency are classified according to whether they must appear in the outcome model, whether they are restricted to consist of past outcomes only, or whether there are no restrictions. These covariates are also classified according to whether they are permitted to be time dependent. Unmeasured (or latent) covariates are classified according to whether they are permitted to be correlated with the outcome process, and whether or not they are permitted to be time dependent. N/A = Not Applicable, † = not all semiparametric joint methods are suitable for this setting (see Table 3 for details), ‡ if correctly specified (has yet to be implemented in practice).

STANDARD METHODS

Generalized Estimating Equations

Mixed models (linear mixed effects or generalized linear mixed effects)

GEE

- GEE requires
  - visits independent of outcomes given covariates in the outcome model
  - $E(\Delta N_i(t) \mid Y_i(t), X_i(t)) = E(\Delta N_i(t) \mid X_i(t))$
- Works for
  - VCAR
  - Special cases of VAR

Mixed effects models

- Linear mixed effects models require
  - Roughly, that gaps between visits are indept of unobserved outcomes given observed outcomes
  - Formally, we require Z_i a vector of auxiliary covariates satisfying

$$f(T_{ik+1} - T_{ik} \mid N(s) : s \le T_{ik}, Y_i(u), X_i(u), Z_i(u) : u \ge 0)$$
$$= f(T_{ik+1} - T_{ik} \mid Y_i^{obs}(s) : s \le T_{ik}, X_i(u), Z_i(u) : u \ge 0)$$
$$E(Y_i(t) \mid X_i(t), Z_i(u) : u \ge 0) = E(Y_i(t) \mid X_i(t))$$

- Works for
  - VCAR
  - Special cases of VAR

INVERSE-INTENSITY WEIGHTING

Inverse-Intensity Weighting

- Similar concept to survey weighting
- Let $\lambda_i(t)$ be the visit intensity for subject i at time t
  - Usually depends on auxiliary covariates $Z_i(t)$
  - Auxiliary covariates may include covariates from the outcome model (Z may contain elements of X)
  - Weight observations by the inverse of visit intensity

Inverse-intensity weighting

- $E(Y_i(t)|X_i(t))$ equal to $E\left(\frac{\Delta N_i(t) Y_i(t)}{\lambda_i(t)} | X_i(t)\right)$

- Fit using software for GEE
  - PROC GENMOD in SAS with scwgt statement
  - geeglm in R with weight argument

Inverse-Intensity Weighting

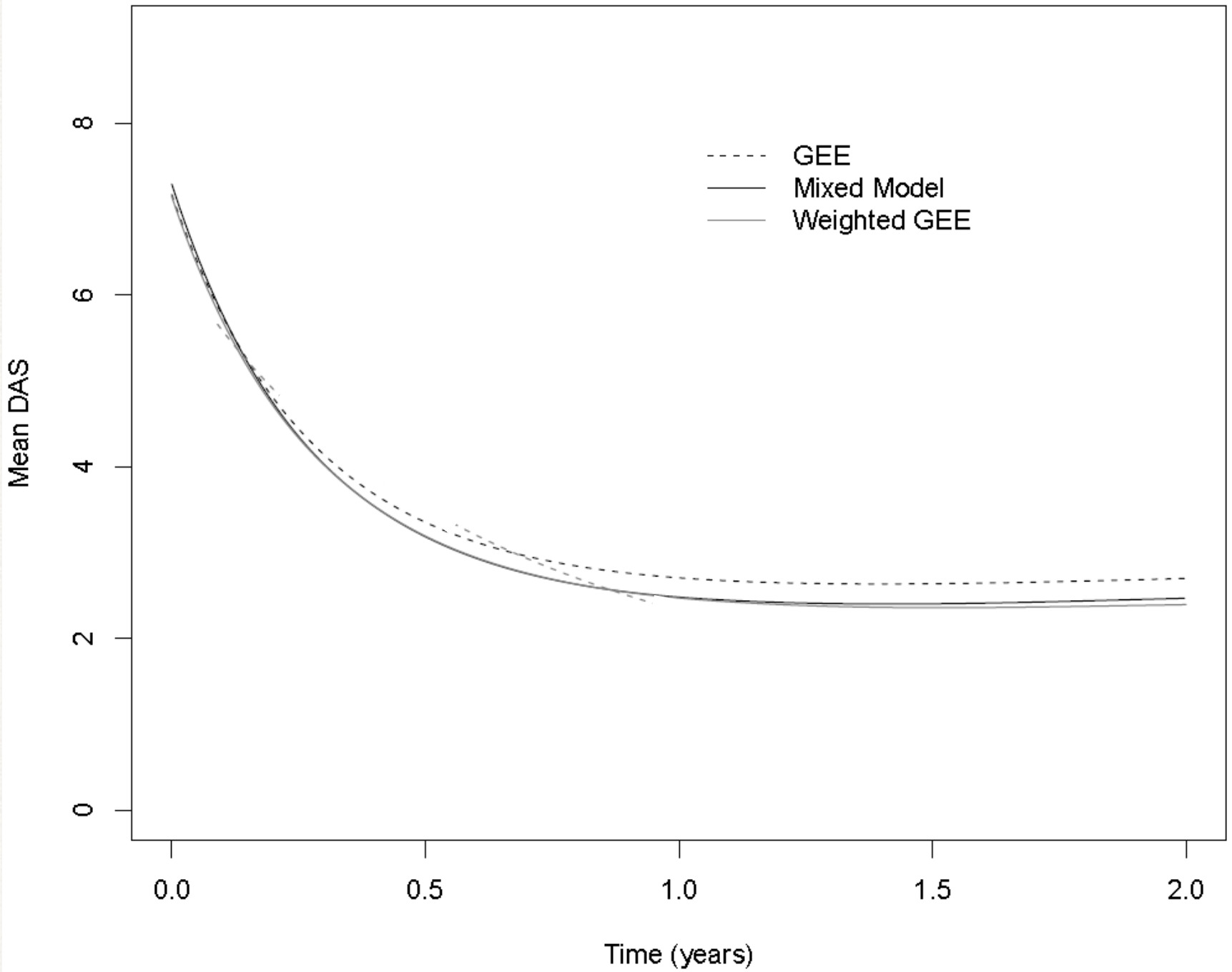
- Require
  - Visits and outcomes to be independent given auxiliary covariates
  - $Y_i(t) \perp\!\!\!\perp N_i(t) \mid Z_i(t)$
- Works for
  - VCAR
  - VAR (any)

JDM Example

- Model visit intensity
  - Disease activity score (DAS) at last visit
  - Non-linear association (non-proportional hazards)
  - No other covariates predictive after accounting for DAS

$$\lambda_i(t) = \lambda_0(t) \exp(0.39 \log(1 + Y_i(T_{N_i(t^-)})))$$

- Weight to be used in analysis is 1/hazard ratio


GEE
Mixed Model
Weighted GEE
Mean DAS
0
2
4
6
8
0.0
0.5
1.0
1.5
2.0
Time (years)

JDM results

Method	AUC	SE
GEE	6.47	0.29
Mixed Model	6.11	0.28
Weighted GEE	6.05	0.25

Multiple Outputation

- Alternative to inverse-intensity weighting
  - Useful in conjunction with analyses for which weighting is problematic
- Opposite of multiple imputation
- Discard observations with probability proportional to $1/\lambda_i(t)$
  - In outputted dataset, visit and outcome processes unassociated
- Analyse, repeat outputation, combine results

SEMI-PARAMETRIC JOINT MODELS

Semi-parametric Joint Models

- Assume that there are latent variables that capture the association between visits and outcomes

Method	Outcome mean model conditional on random effects	Intensity model for N_i^* conditional on outcomes and random effects
Liang et al.[6]	$\beta_0(t) + \mathbf{X}_i(t)\boldsymbol{\beta} + \mathbf{W}_i(t)\mathbf{V}_{i1}$	$V_{i2}\lambda_0(t)\exp(\mathbf{Z}_i\gamma)$
Sun et al.[12]	$\beta_0(t; V_{i1}) + \mathbf{X}_i\boldsymbol{\beta}$	$\lambda_0(t; V_{i2})\exp(\mathbf{X}_i\gamma)$
Sun et al.[8]	$V_i\beta_0(t)\exp(\mathbf{X}_i\boldsymbol{\beta})$	$V_i\lambda_0(t)\exp(\mathbf{X}_i\gamma)$
Song et al.[11]	$\beta_0(t) + \mathbf{X}_i(t)\boldsymbol{\beta} + V_{i1}$	$V_{i2}\lambda_0(t)\exp(\mathbf{X}_i(t)\gamma)$

Here $\mathbf{X}_i$ is a row vector of covariates, which when denoted by $\mathbf{X}_i(t)$ is permitted to be time dependent. V_i, V_{i1}, and V_{i2} are random effects, $\mathbf{W}_i(t)$ is a subset of the covariates $\mathbf{X}_i(t)$, $\mathbf{Z}_i$ is a vector of auxiliary baseline covariates, C_i is a censoring time, λ_0 is a baseline hazard, and $\beta_0, \boldsymbol{\beta}, \gamma$ are regression coefficients. The intensity model for $N_i^*(t)$ is conditional on $\bar{Y}_i(\infty)$ as well as any random effects and covariates.

Semi-parametric Joint Models

- Currently no ready-made code to implement
- Work with
  - VCAR
  - Some cases of VAR
  - Some cases of VNAR

Example

- RCT of thiotepa vs. placebo for bladder cancer
- f/u scheduled for every 3 months
- Number of new tumours noted
- Much deviation in visit times
  - Likely patient-driven
- Target of inference: effect of thiotepa on rate of new tumours

$$E(Y_i(t)|X_i) = \beta_0(t)\exp(X_i\beta)$$

Visit process

- Measured covariates not predictive of visit intensity

Covariate	Hazard Ratio	P-value
Group	1.07	0.6
# new tumours at last visit	0.97	0.27
Any new tumours at last visit	0.97	0.86

- VCAR or VNAR
- VNAR more plausible

Semi-parametric Joint Models

- Assume that there are latent variables that capture the association between visits and outcomes

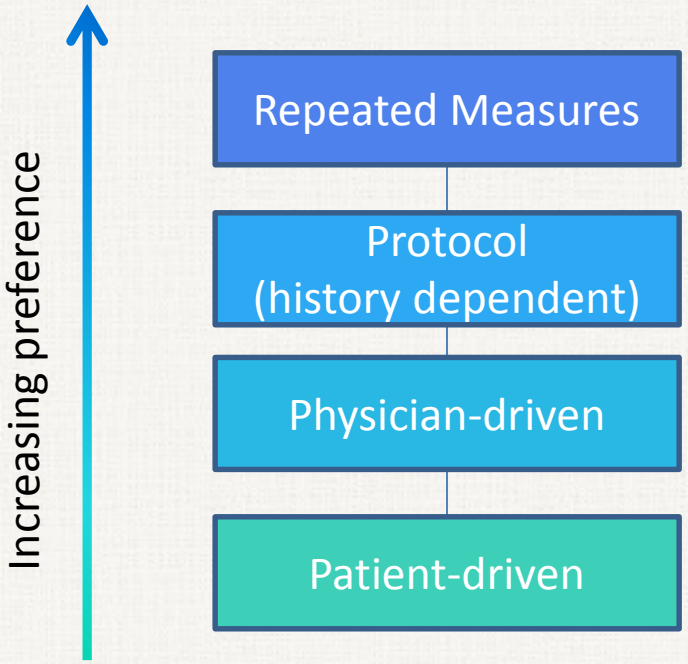
Method	Outcome mean model conditional on random effects	Intensity model for N_i^* conditional on outcomes and random effects
Liang et al.[6]	$\beta_0(t) + \mathbf{X}_i(t)\boldsymbol{\beta} + \mathbf{W}_i(t)\mathbf{V}_{i1}$	$V_{i2}\lambda_0(t)\exp(\mathbf{Z}_i\gamma)$
Sun et al.[12]	$\beta_0(t; V_{i1}) + \mathbf{X}_i\boldsymbol{\beta}$	$\lambda_0(t; V_{i2})\exp(\mathbf{X}_i\gamma)$
Sun et al.[8]	$V_i\beta_0(t)\exp(\mathbf{X}_i\boldsymbol{\beta})$	$V_i\lambda_0(t)\exp(\mathbf{X}_i\gamma)$
Song et al.[11]	$\beta_0(t) + \mathbf{X}_i(t)\boldsymbol{\beta} + V_{i1}$	$V_{i2}\lambda_0(t)\exp(\mathbf{X}_i(t)\gamma)$

Here $\mathbf{X}_i$ is a row vector of covariates, which when denoted by $\mathbf{X}_i(t)$ is permitted to be time dependent. V_i, V_{i1}, and V_{i2} are random effects, $\mathbf{W}_i(t)$ is a subset of the covariates $\mathbf{X}_i(t)$, $\mathbf{Z}_i$ is a vector of auxiliary baseline covariates, C_i is a censoring time, λ_0 is a baseline hazard, and $\beta_0, \boldsymbol{\beta}, \gamma$ are regression coefficients. The intensity model for $N_i^*(t)$ is conditional on $\bar{Y}_i(\infty)$ as well as any random effects and covariates.

Bladder RCT Results

Method	Rate Ratio	95% CI
GEE (log link)	0.47	0.21 to 1.02
Sun semi-parametric joint model	0.49	0.24 to 1.01

Could we design studies better?



- Fixed visits
- Have a protocol
- Record recommended visit intervals
- Keep track of deviations
- Request minimal follow-up

Summary

- Irregular visiting can result in bias
- Need to consider why visits are irregular
- Use an appropriate analysis
- Be thoughtful about study design

Implementation

- R package that:
  - Calculates inverse-intensity weights
  - Fits an inverse-intensity weighted GEE
  - Produces a multiply outputted dataset
  - Performs multiple outputation for any given analytic technique (GEE, LME, etc.)

- Email me if interested in being a tester!

Learning objectives

- Explain why irregular observation can be a problem
- Understand different visiting patterns
- Choose an appropriate analysis for a given visiting pattern

Key References

- Pullenayegum EM, Lim L. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. Statistical Methods in Medical Research (ePub ahead of print)
- Lin H, Scharfstein D, Rosenheck R. Analysis of longitudinal data with irregular, outcome-dependent follow-up. Journal of the Royal Statistical Society, Series B 2004;66:791-813.
- Follmann D, Proschan M, Leifer E. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. Biometrics 2003; 59:420-429.
- Pullenayegum EM. Multiple outputation for the analysis of longitudinal data subject to irregular observation. Statistics in Medicine (in press)