Visualising and analysing time-to-event data: lifting the veil of censoring

Patrick Royston

Cancer Group, MRC Clinical Trials Unit, London

A poet writes about censored observations:

Last night I saw upon the stair
A little man who wasn't there
He wasn't there again today
Oh, how I wish he'd go away!

From *Antigonish* (1899)

Hughes Mearns (1875-1965)

Outline

- Why is censoring of time-to-event data an issue?
- Example in breast cancer
- Visualisation of censored data using model-based imputation
- Multiple imputation and analysis of survival data with missing covariate observations
- Demonstration with Stata

Why is censoring an issue?

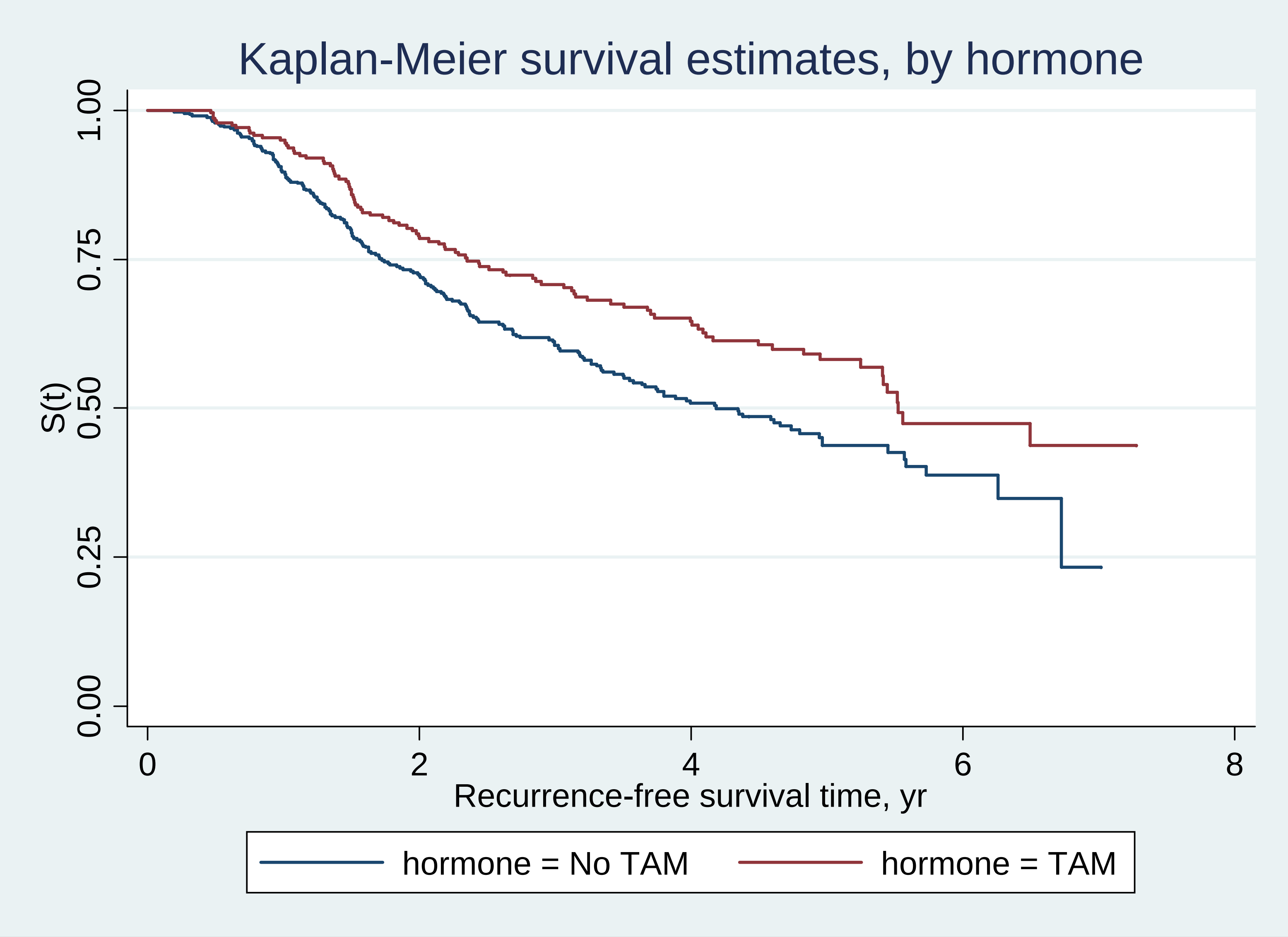
- You can't picture the raw data easily
- Reliance on Kaplan-Meier plots
  - Exaggerates differences between groups
  - Attracts attention to unreliable survival estimates at extreme times
- Data will be analysed using Cox model
  - Still the almost-automatic choice – although decent alternatives exist
- Time is "forgotten about" in the Cox model
  - Analysis is based on the ranks of failure times

- Results of Cox regression models are usually expressed as (log) hazard ratios
  - Indirect – not dealing directly with time
  - Can be hard to interpret – different effect on survival curves at high and low survival probs
  - Particularly difficult for interactions – ‘ratio of hazard ratios’
- Non-proportional hazards
  - Data with long-term follow-up typically have it
  - Modelling and interpretation may be complex

Example: Primary node-positive breast cancer

- GBSG trial BMFT-2
- 686 patients, 299 events for recurrence-free survival (RFS)
- Patients assigned to hormonal therapy (TAM) or not
- Visualise the effect of TAM on RFS
- Visualise interaction between TAM and ER (estrogen receptor status)

Traditional visualisation: Kaplan-Meier by TAM group



Dot plot by TAM therapy – unhelpful with censored data

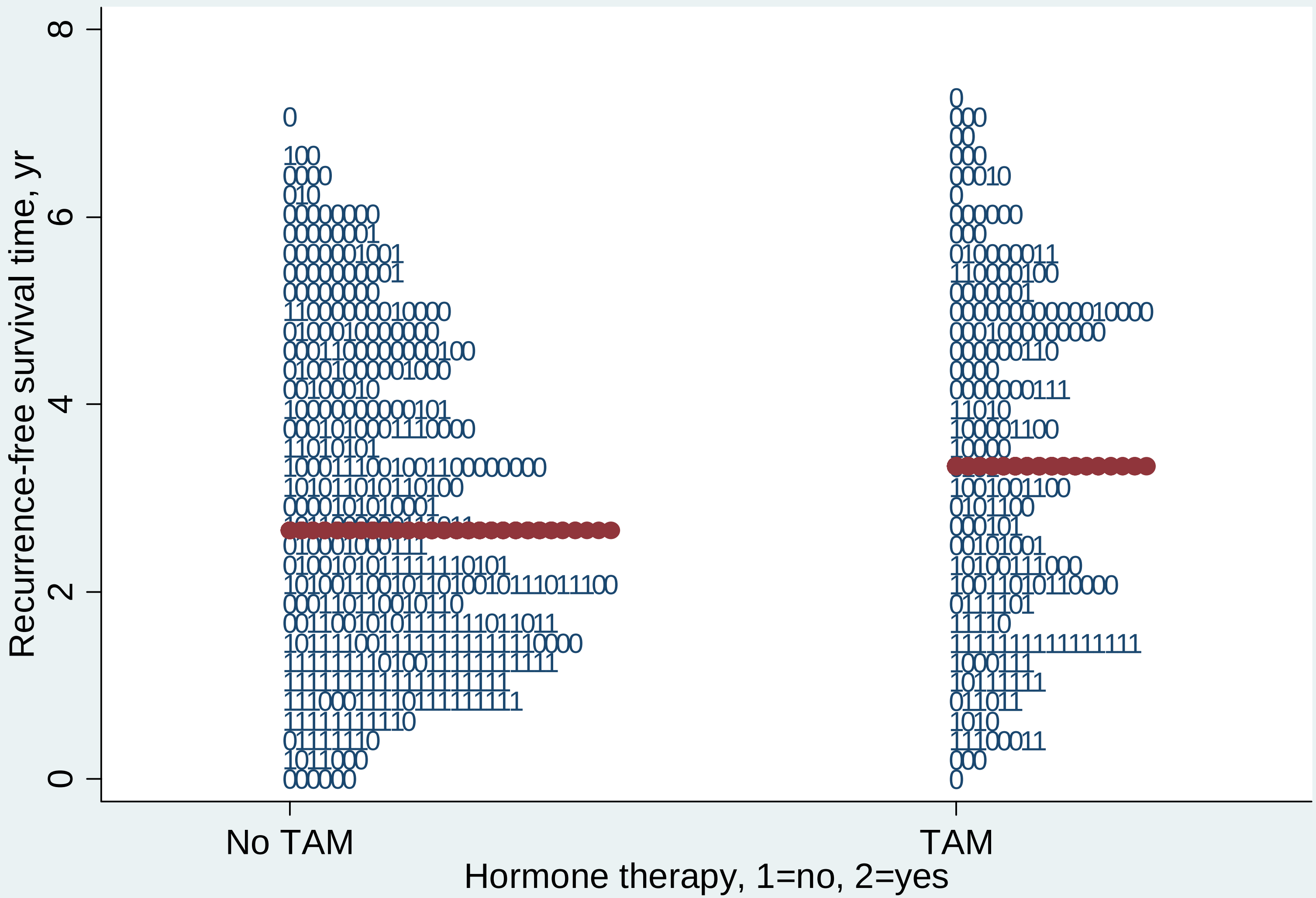


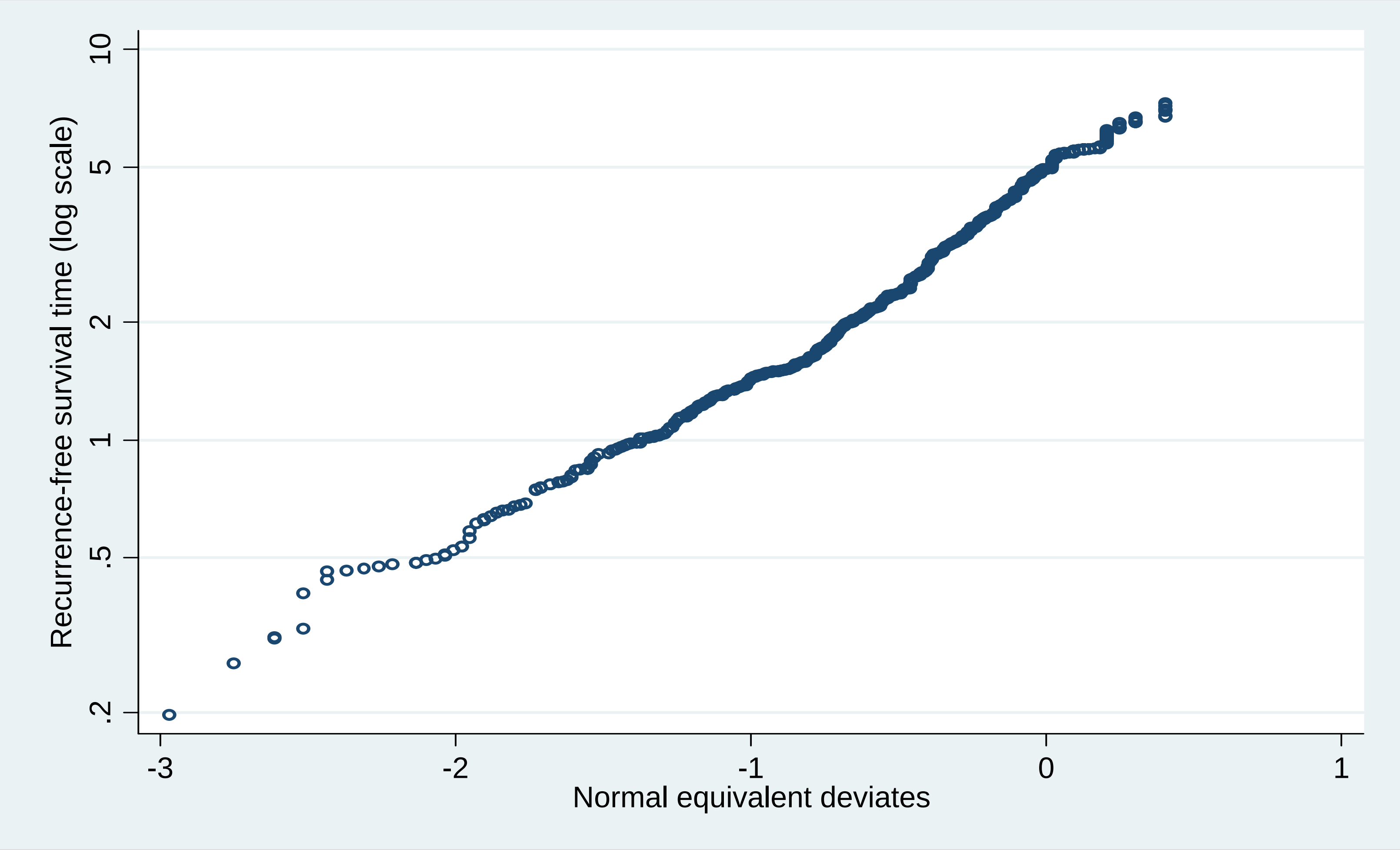
fig2

How better to visualise survival times?

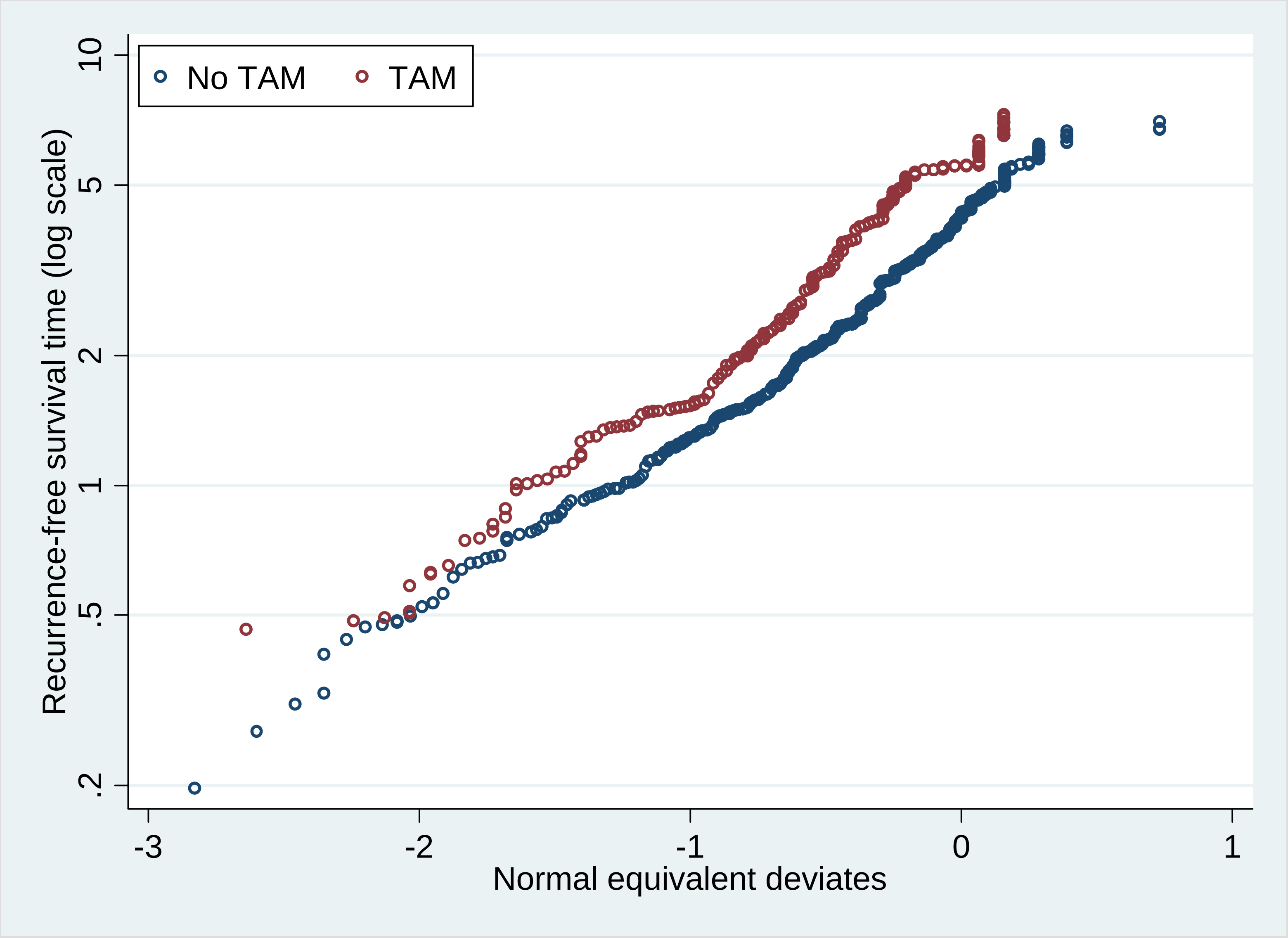
- To make progress with visualisation, aim to impute the “missing” part of censored times
- Assume a parametric distribution of survival time
- Survival times are sometimes approximately lognormally distributed (Royston 2001a)
  - Can check by using modified Normal Q-Q plot
- If lognormal approximation is not good, can consider Box-Cox transformation of time
  - Or another transformation towards normality

Assessing lognormality: modified Normal Q-Q plot

- Simple transformation of Kaplan-Meier survival curve



Normal Q-Q plot by TAM group



Visualisation of censored data using imputation

- Create m (≥ 1) copies of the data with censored survival times imputed
- Need an imputation model to reflect
  - Distribution of times (e.g. lognormal)
  - Effects of covariates (prognostic factors)
- Creating an imputation model:
  - Use `mfp` with cnreg (censored normal regrn.) to model poss. non-linear effects of covariates
  - E.g. `mfp cnreg lnt x1 x2 x3 x4a x4b x5 x6 x7 hormone, censored(c) select(1) dfdefault(2)`

Creating the imputed dataset(s)

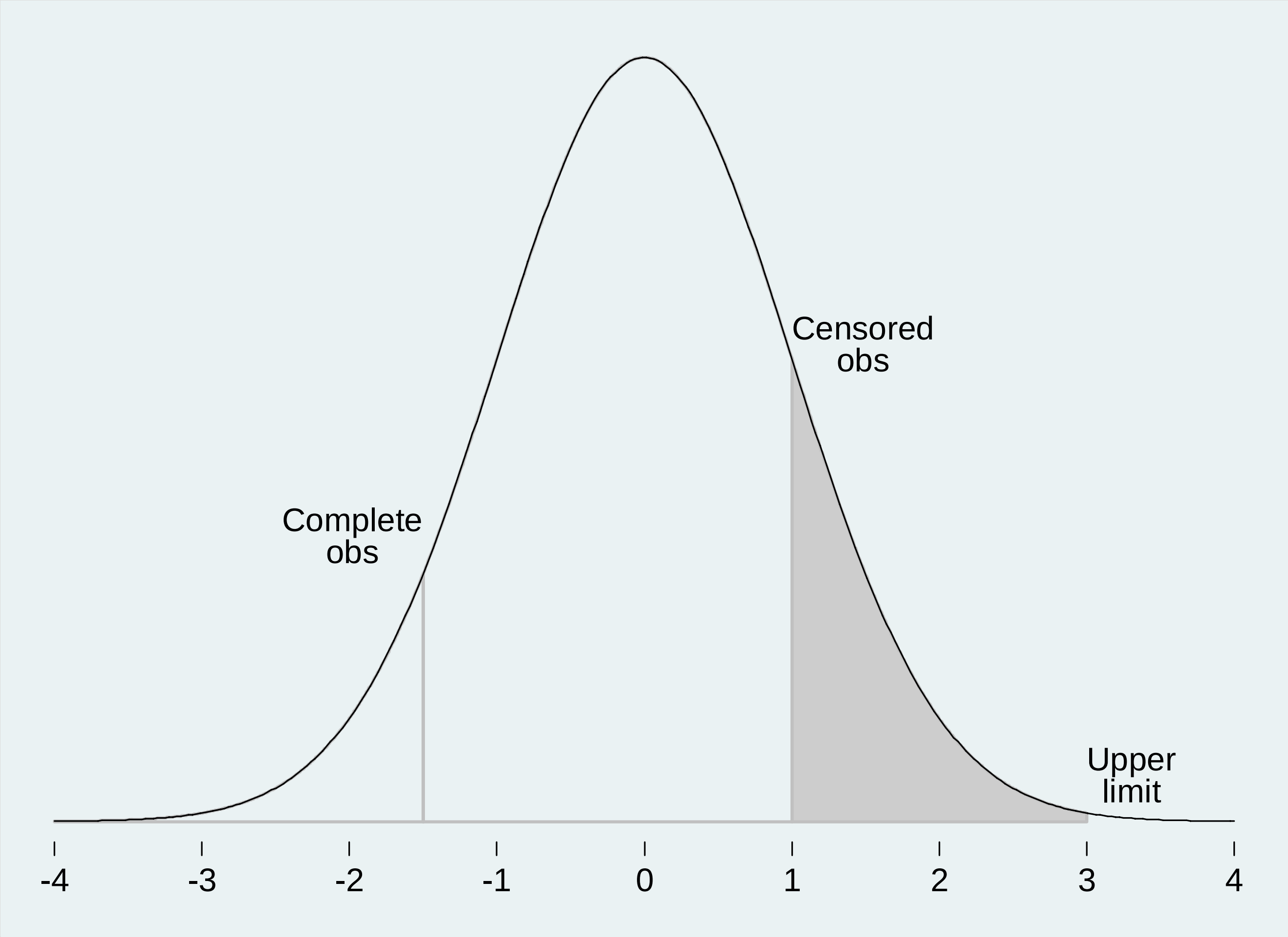
- Can use the **`ice`** multiple imputation command to create the imputations
  - Royston (2004, 2005a, 2005b) *Stata J*
- **ice** ***varlist*** **using** ***filename*[.dta] [if *exp*] [in *range*] [*weight*], [m(#) cmd(*cmdlist*) cycles(#) boot[(*varlist*)] seed(#) dryrun eq(*eqlist*) passive(*passivelist*) substitute(*sublist*) dropmissing interval(*intlist*) *other_options*]**

Interval censoring with `ice`

- `gen ll = lnt`
- `gen ul = cond(_d==1, lnt, ln(50))`
  `// chose upper limit of 50 years for RFS: can use . for +∞`
- (generate FP transformations of `x1`, `x5`, `x6`)
- `ice x1_1 x2 x3 x4a x4b x5_1 x6_1 x7 hormone ll ul lnt using imputed.dta, interval(lnt:ll ul) m(10)`

How `interval()` works

- Sample randomly from truncated normal distribution (shaded)



Code fragment from uvis.ado

MRC Clinical Trials Unit

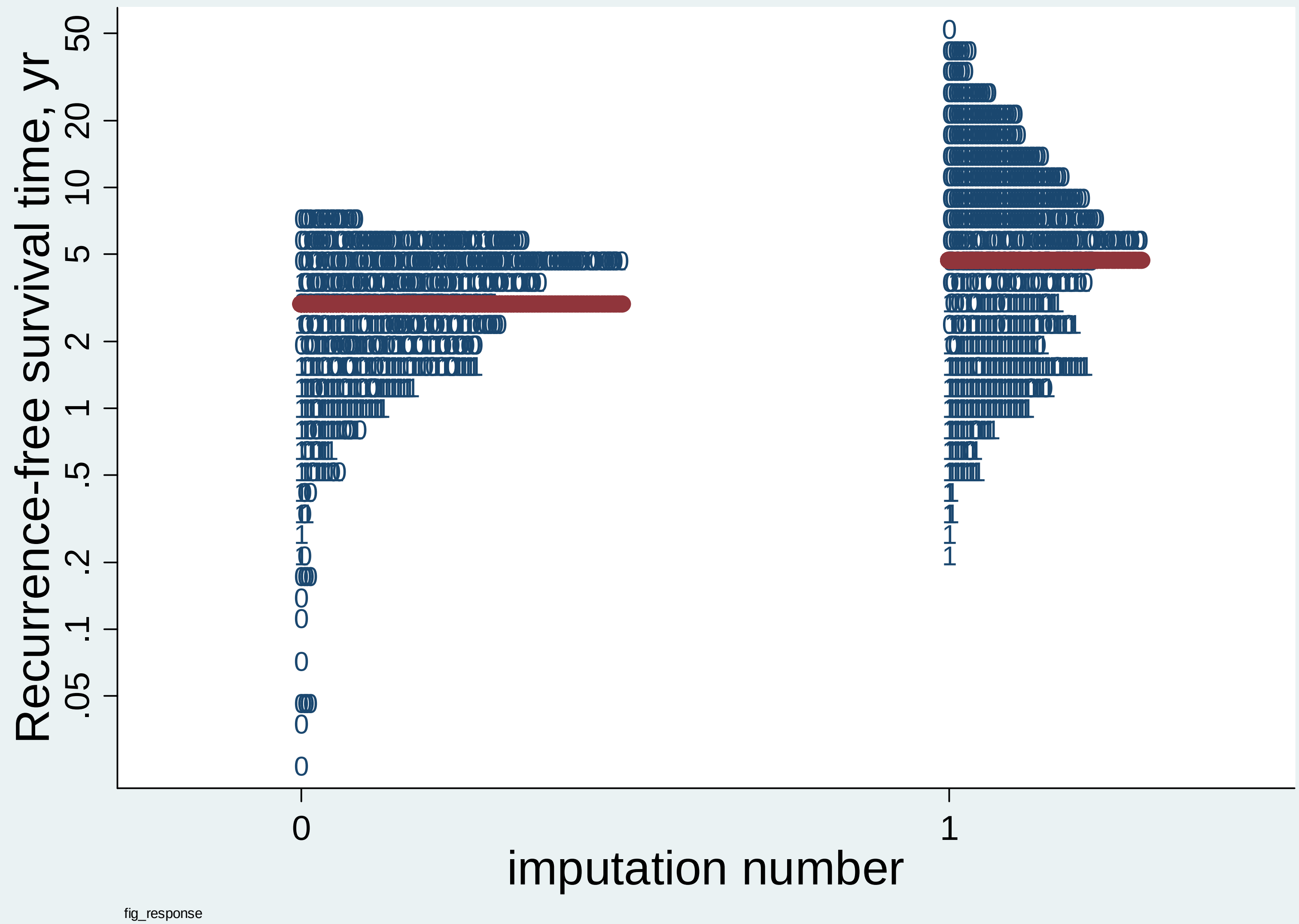
```
`cmd' `yvarlist' `xvars' `wgt',
  `options'
...
if "`cmd'"=="intreg" {
    tempvar PhiA PhiB
    gen `PhiA‘ = cond(missing(`ll'), 0,
      norm((`ll'-`xb')/`rmsestar'))
    gen `PhiB‘ = cond(missing(`ul'), 1,
      norm((`ul'-`xb')/`rmsestar'))
    replace `yimp‘ = `xb‘
      +`rmsestar'*invnorm(`u'*
      (`PhiB'-`PhiA')+`PhiA')
}
```

Uses of the `interval()` option

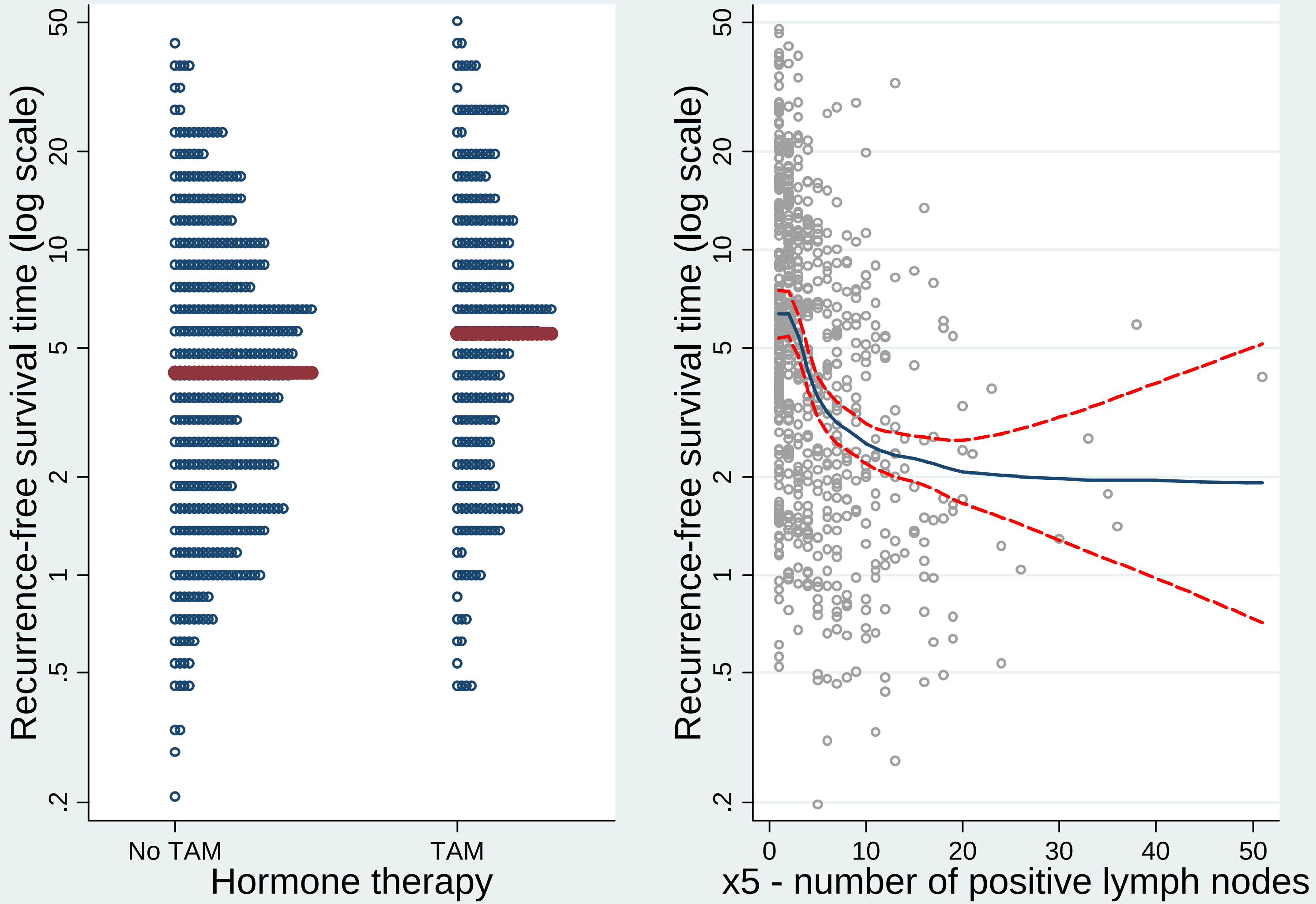
- Impute right-, left- or interval-censored outcomes
  - Response variable in time-to-event studies
- Impute when a covariate is sometimes partly observed, sometimes complete
  - Some observations recorded exactly
  - Others known to be below or above a cutoff
  - E.g. D-dimer in DVT, PgR/ER in breast cancer
- Interval censored covariates
  - Income in surveys recorded as ranges only

Breast cancer data: visualisation of time to recurrence

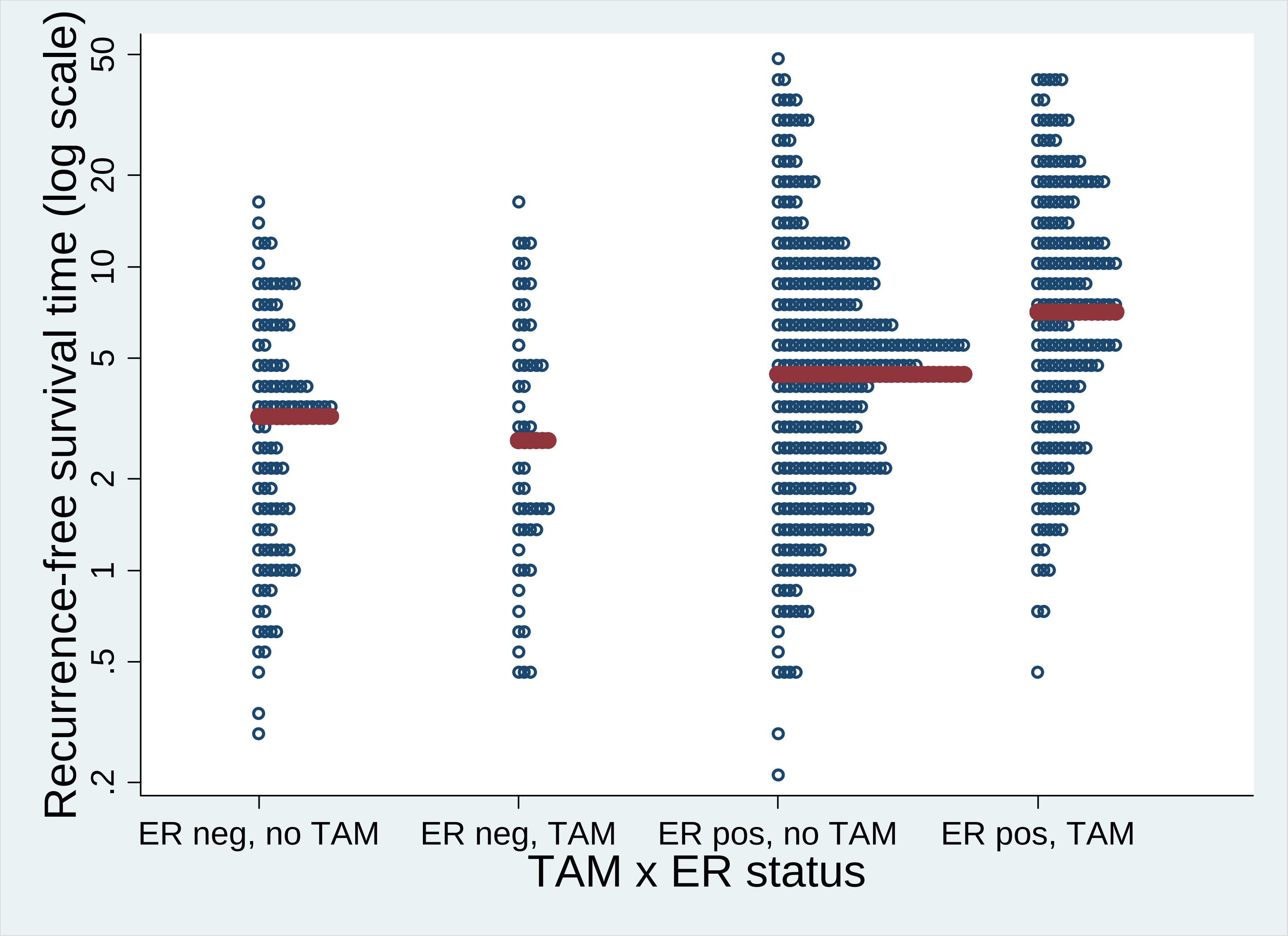
MRC Clinical Trials Unit



Visualisation: some plots using the first imputed sample



Visualisation: treatment by covariate interaction



Limitations

MRC Clinical Trials Unit

- Imputed times to event are helpful for visualisation, but less so for analysis
  - Effectively, such imputations are extrapolations into the future
  - We don't **know** the future distribution
  - Estimates of means, SD's, regression coeffs etc. are heavily dependent on the distributional assumptions
  - Potential for bias if assumed distr'n is wrong
- Imputed times may be unrealistic
  - E.g. survival time 150 years!

Other approaches

- A reasonably large literature exists
- Buckley-James estimation (Buckley & James 1979)
  - Estimates the mean of the censored part
  - Not so good for visualisation
- Wei & Tanner (1991)
  - Two algorithms which give multiple imputations of the censored part
  - Relaxes the normality assumption – samples taken from the distribution of the residuals
- **`stpm`** (Royston 2001b, Royston & Parmar 2002)
  - More flexible distributions of survival time available

Imputation of survival data with missing covariate observations

- So far, have assumed covariates have complete data
- If covariates have **missing data**, need a suitable algorithm for multiple imputation of all missing values
  - e.g. MICE (`ice`)
- To reduce bias, must include the response (time-to-event) in the imputation model
  - How?
- "Standard" approach is to include (censored) **log time** and the **censoring indicator** in the imputation model
  - No theoretical justification
- May be better to
  - Include covariates as usual
  - Impute right-censored times using `ice` with `interval()` option
- Can also use imputed data for visualisation

Analysis of survival data with missing covariate observations

- Disregard the imputed times in the MI dataset
  - Except for visualisation purposes
- Use original time and censoring indicator
- Can analyse the MI dataset using
  - `stcox` (Cox regression)
  - `streg` (several models available)
  - `stpm` (flexible parametric survival models)
- `micombine` supports such models

Conclusions

- Use of familiar graphical tools with imputed times to event can give greater insight into censored survival data
  - Scatter plots, smoothers, etc
- Treatment or prognostic effects may be depressingly small when displayed as scatter plots of times
  - Much overlap between groups
  - Weak regression relationships
- Imputation of times may be helpful in multiple imputation with missing covariate values

Some references

- Buckley J, James I (1979) Linear regression with censored data. *Biometrika* **66**: 429-436
- Faucett CL, Schenker N, Taylor JMG (2002) Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* **58**: 37-47.
- Ma SG (2006) Multiple augmentation with partial missing regressors. *Biometrical Journal* **48**: 83-92
- Pan W (2000) A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**: 199-203
- Royston P (2001a) The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica* **55**: 89-104
- Royston P (2001b) Flexible alternatives to the Cox model, and more. *The Stata Journal* **1**:1-28.
- Royston P, Parmar MKB (2002) Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* **21**: 2175-2197
- Royston P (2004) Multiple imputation of missing values. *Stata Journal* **4**: 227-241
- Royston P (2005a) Multiple imputation of missing values: update. *Stata Journal* **5**: 188-201
- Royston P. (2005b) Multiple imputation of missing values: update of ice. *Stata Journal* **5**: 527-536
- Tanner MA, Wing HW (1987) The calculation of posterior distributions by data augmentation. *JASA* **82**: 528-540. [Cited 642 times, WoS 10.9.2006]
- Wei GCG, Tanner MA (1990) Posterior computations for censored regression data. *JASA* **85**: 829-39
- Wei GCG, Tanner MA (1991) Application of multiple imputation to the analysis of censored regression data. *Biometrics* **47**: 1297-1309