# Hierarchical Survival Models: Estimation, Prediction, Interpretation

Alessandro Gasparini      Red Door Analytics AB

2023 Northern European Stata Conference

2023-09-01

# Hierarchical

Hierarchical (or multilevel) data can occur frequently in a wide range of research disciplines:

- In education, students can be clustered within schools and districts;

- In longitudinal studies, repeated observations can be nested within study participants;

- In medicine, patients can be clustered within hospitals, and then within regions;

- Etc.

The units of analysis are usually individuals or observations (at the lowest level) who are nested within contextual (or aggregate) units (at a higher level).

# Hierarchical Survival

Analogously, hierarchical survival data arise when we observe time to event data, recorded at the lowest level and nested within one (or more) higher level units.

Examples, from the field of medicine:

- Time to death in patients with cardiovascular disease nested within surgeons and then within hospitals;

- Time to blindness in patients with diabetic retinopathy, where both eyes are nested within each subject;

- Individual-patient-data (IPD) meta-analysis, where subjects are nested within the different studies being combined;

- Etc.

# Hierarchical Survival Models

Several methods have been proposed throughout the years to analyse hierarchical survival data.

For the purposes of this presentation, we will focus on mixed-effects, proportional-hazards survival regression models. These can be fit in Stata using a variety of built-in commands (such as `mestreg` and `gsem`) and user-contributed commands (such as `stmixed` and `merlin`).

One of the benefits of mixed-effects models is that they can be used to obtain both subject-specific and population-level effects. Some other methods can only do the latter.

# Hierarchical Survival Models: Notation

Assuming a two-levels nesting structure (patients nested within hospitals), a mixed-effects proportional hazards survival model can be defined as:

$$h_{ij}(t) = h_0(t) \exp(X_{ij}\beta + Z_{ij}b_j)$$

where $i$ indexes patients and $j$ indexes hospitals.

$h_0(t)$ denotes a baseline hazard function.

$X_{ij}$ denotes a set of observed covariates with associated regression coefficients $\beta$, often named *fixed effects* in medical statistics.

$Z_{ij}$ denotes a set of $k$ latent variables with associated *random effects* $b_j$. $X_{ij}$ and $Z_{ij}$ can overlap, in principle.

Finally, the $k$ random effects $b_j$ are assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\Sigma_b$: $b_j \sim N_k(0, \Sigma_b)$.

The following model is often referred to as a *random intercept* model:

$$h_{ij}(t) = h_0(t) \exp(X_{ij}\beta + b_j)$$

In this case, the only random effect, $b_j$, follows a univariate normal distribution.

This formulation is close to that of shared frailty survival models,

$$h_{ij}(t) = h_0(t) \exp(X_{ij}\beta)\alpha_j,$$

where the frailty term, $\alpha_j$, acts multiplicatively: intuitively, $\exp(b_j) = \alpha_j$. Common distributions for the frailty term are Gamma, log-Normal, etc. Shared frailty survival models can be fit in Stata using built-in commands such as `stcox` and `streg`.

# Case Study

We illustrate hierarchical survival models in practice using a dataset that combines data from several clinical trials in advanced ovarian cancer:

```
. list if patientID <= 5

   +----------------------------------------------------------------------+
   | patien~D   trialID   trt        timeS   statusS        timeT   statusT |
   |----------------------------------------------------------------------|
1. |        1         2     0    .10515873         1    .18571429         1 |
2. |        2         2     0    .89523809         1    1.4087302         1 |
3. |        3         2     0    .07896825         1    .12619048         1 |
4. |        4         2     1    1.7392857         0    1.7392857         0 |
5. |        5         2     0    .09126984         1    .12738095         1 |
   +----------------------------------------------------------------------+
```

Specifically, here we aim to study the effect of a certain treatment (`trt`) on progression-free survival (`timeS`, `statusS`) while accounting for clustering within different trials (`trialID`).

## stset Data

Given that we are working with survival data, we start with the usual stset command:

```
. stset timeS, failure(statusS == 1)

Survival-time data settings

         Failure event: statusS==1
Observed time interval: (0, timeS]
     Exit on or before: failure
--------------------------------------------------------------------------
      1,192  total observations
          0  exclusions
--------------------------------------------------------------------------
      1,192  observations remaining, representing
        977  failures in single-record/single-failure data
    543.942  total analysis time at risk and under observation
                                             At risk from t =           0
                                  Earliest observed entry t =           0
                                       Last observed exit t =    1.983333
```

## Modelling

Then, we fit a random-intercept hierarchical survival model:

$$h_{ij}(t) = h_0(t) \exp(\text{trt}_{ij}\beta + b_j)$$

for a certain $i^{\text{th}}$ subject from the $j^{\text{th}}$ trial. For this, we use the `mestreg` command and test both an exponential and Weibull survival distribution for the baseline hazard function $h_0$:

```
. mestreg trt || trialID:, dist(exp)
[ Output omitted... ]
. estimates store me_rint_exp
```

```
. mestreg trt || trialID:, dist(wei)
[ Output omitted... ]
. estimates store me_rint_wei
```

Using the `stmixed` command we could use/test additional parametric (such as the Gompertz) or flexible parametric formulations for $h_0$.

# Best-Fitting Model

The can then use `estimates stats` to calculate the AIC and BIC:

```
. estimates stats *

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
      Model |        N    ll(null)  ll(model)     df         AIC         BIC
------------+----------------------------------------------------------------
 me_rint_exp |     1,192          .  -378.7514      3    763.5028    778.7529
 me_rint_wei |     1,192          .  -310.2967      4    628.5935     648.927
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] IC note.
```

For both information criteria, *the lower the better fit*. In this case, the Weibull model seems to provide the better fit between the two, and we will be assuming so from now onwards.

# Random Treatment Effect

For comparison purposes, we also fit the following random intercept and treatment model:

$$h_{ij}(t) = h_0(t) \exp[\mathtt{trt}_{ij}(\beta + b_{1j}) + b_{0j}]$$

In Stata:

```
. mestreg trt || trialID: trt, dist(wei) nohr cov(unstr)
. estimates store me_rboth_wei
```

This however did not fit significantly better, so we stick with the random intercept model:

```
. lrtest me_rboth_wei me_rint_wei

Likelihood-ratio test
Assumption: me_rint_wei nested within me_rboth_wei

 LR chi2(2) =    0.55
Prob > chi2 = 0.7591
```

```
. estimates restore me_rint_wei
. mestreg, notable

Mixed-effects Weibull PH regression            Number of obs     =       1,192
Group variable: trialID                        Number of groups  =          50

                                               Obs per group:
                                                            min =           2
                                                            avg =        23.8
                                                            max =         274

Integration method: mvaghermite                Integration pts.  =           7

                                               Wald chi2(1)      =       12.27
Log likelihood = -310.29674                    Prob > chi2       =      0.0005
LR test vs. Weibull model: chibar2(01) = 18.97      Prob >= chibar2 = 0.0000
```

# Modelling Results (2)

```
. mestreg, noheader nohr

------------------------------------------------------------------------------
         _t | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
------------+-----------------------------------------------------------------
        trt | -.2256597   .0644333    -3.50   0.000    -.3519467   -.0993727
      _cons |  .5363975    .062564     8.57   0.000     .4137743    .6590208
------------+-----------------------------------------------------------------
       /ln_p | -.2775082   .025443                     -.3273755   -.2276409
------------+-----------------------------------------------------------------
trialID     |
   var(_cons)|  .0470106   .0234363                     .0176947    .1248956
------------------------------------------------------------------------------
LR test vs. Weibull model: chibar2(01) = 18.97       Prob >= chibar2 = 0.0000
```

# Modelling Results (3)

```
. mestreg, noheader

--------------------------------------------------------------------------------
         _t |  Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
------------+-------------------------------------------------------------------
        trt |   .7979896    .0514171    -3.50   0.000     .7033176    .9054052
      _cons |   1.709836    .1069742     8.57   0.000     1.512516    1.932899
------------+-------------------------------------------------------------------
       /ln_p |  -.2775082    .025443                      -.3273755   -.2276409
------------+-------------------------------------------------------------------
trialID     |
   var(_cons)|   .0470106    .0234363                      .0176947    .1248956
--------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
Note: _cons estimates baseline hazard (conditional on zero random effects).
LR test vs. Weibull model: chibar2(01) = 18.97        Prob >= chibar2 = 0.0000
```

# Modelling Results: Interpretation

The estimated hazard ratio for treatment was 0.798 (95% C.I.: 0.703, 0.905). We interpret this as treated subjects having a hazard of event that is about 80% of that of untreated patients.

Note that this hazard ratio must be interpreted as a *conditional (on the random effects) hazard ratio* (that is, for a certain trial/cluster) and that proportional hazards only holds conditionally.

At a population level, marginal hazards from a proportional hazards model are generally not proportional.

The likelihood ratio test vs a model without random intercept supports the model with random effects (Prob >= chibar2 = 0.0000).

_cons and \ln_p are the estimated (log-) coefficients for the baseline hazard Weibull distribution (scale and shape, respectively).

var(_cons) is the estimated variance of the normal distribution that we assumed for the random intercept. We will get back to this in a few slides.

# Predictions: Conditional vs Marginal

After fitting a survival model, we are usually interested in predicting either the hazard or the survival function (more prediction types are of course possible). In the settings of hierarchical models, however, we need to distinguish between two prediction types: *conditional* or *marginal* (w.r.t. the random effects).

Specifically:

- Conditional predictions require calculating the group-specific random effects, to combine with the estimated fixed effects (i.e., the regression coefficients and ancillary parameters). Sometimes we can obtain predictions for fixed values of the random effects, e.g., by setting them to zero;

- Marginal predictions are obtained by integrating out (i.e., *marginalising over*) the random effects distribution. These are often referred to as *population-level* predictions.

```
. replace _t = 2
(1,192 real changes made)

. predict S_cond, surv conditional
. predict S_fixed, surv conditional(fixedonly)
. predict S_marg, surv marginal

. list patientID trialID trt S_* ///
>         if patientID == 1 | patientID == 4 | patientID == 129 | patientID == 130

      +---------------------------------------------------------+
      | patien~D   trialID    trt     S_cond    S_fixed    S_marg |
      |---------------------------------------------------------|
   1. |       1         2      0   .3446461   .0555253   .0623801 |
   4. |       4         2      1   .4273943   .0995672   .1063518 |
 129. |     129         1      0   .3370368   .0555253   .0623801 |
 130. |     130         1      1   .4198474   .0995672   .1063518 |
      +---------------------------------------------------------+
```
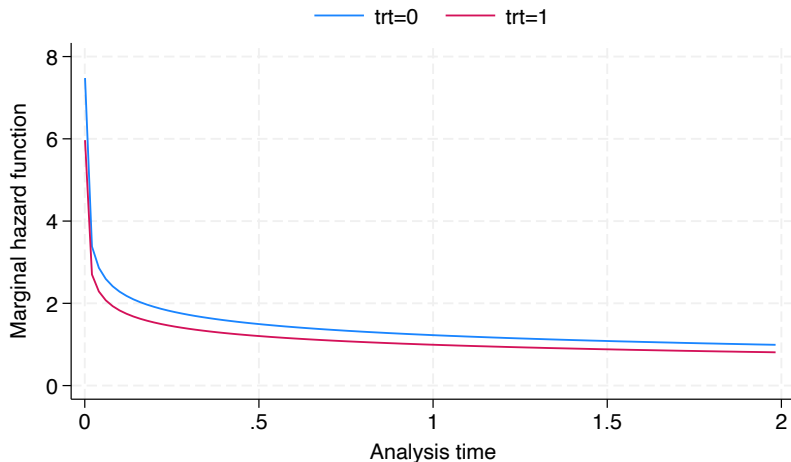
```
. stcurve, hazard at1(trt=0) at2(trt=1) fixedonly ///
>          title("") legend(rows(1) position(12))
```
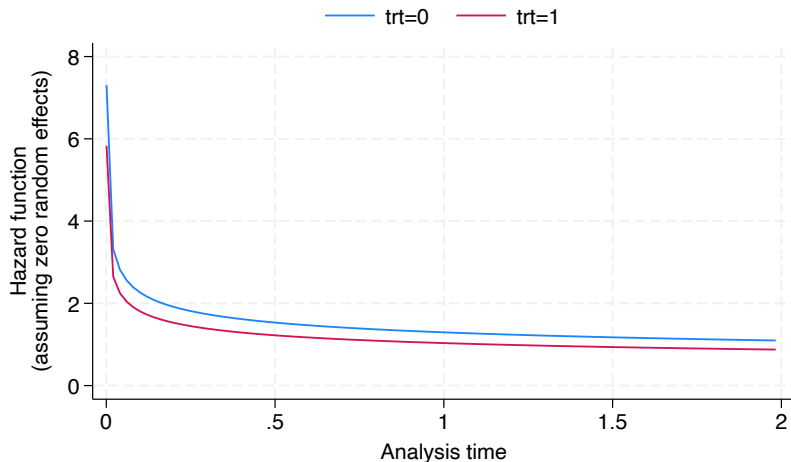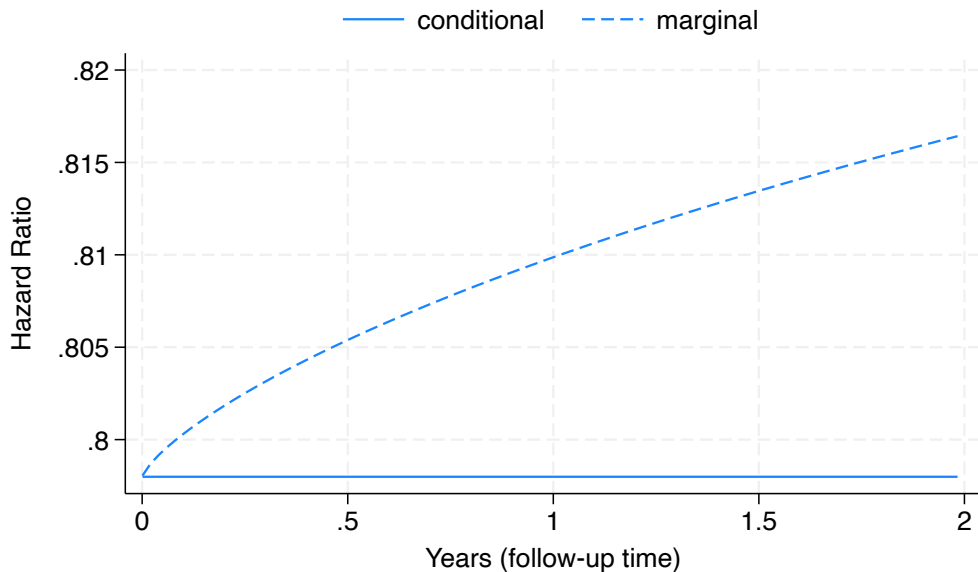
```
. stcurve, hazard at1(trt=0) at2(trt=1) marginal ///
>         title("") legend(rows(1) position(12))
```

# Predictions: Hazard Ratios
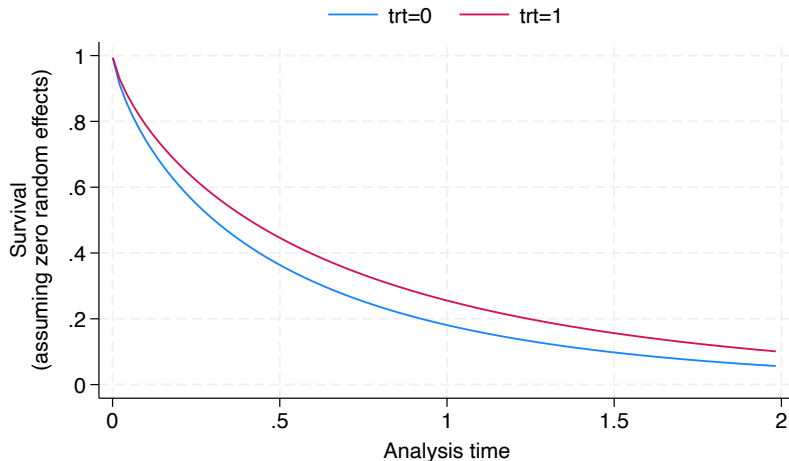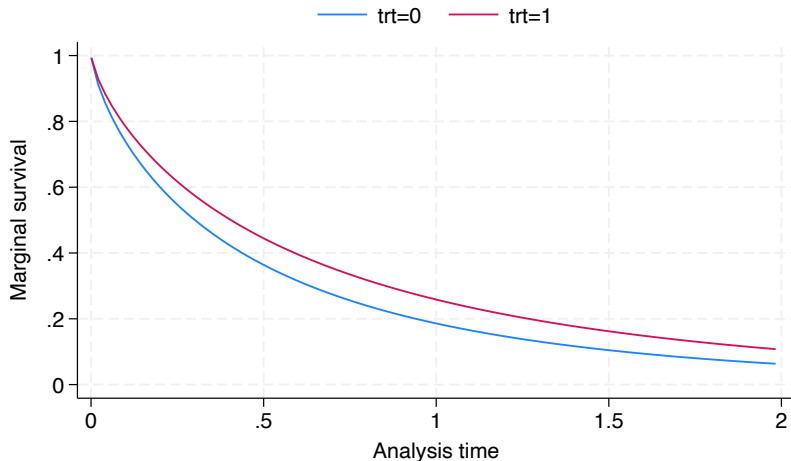
# Predictions: Conditional Survival

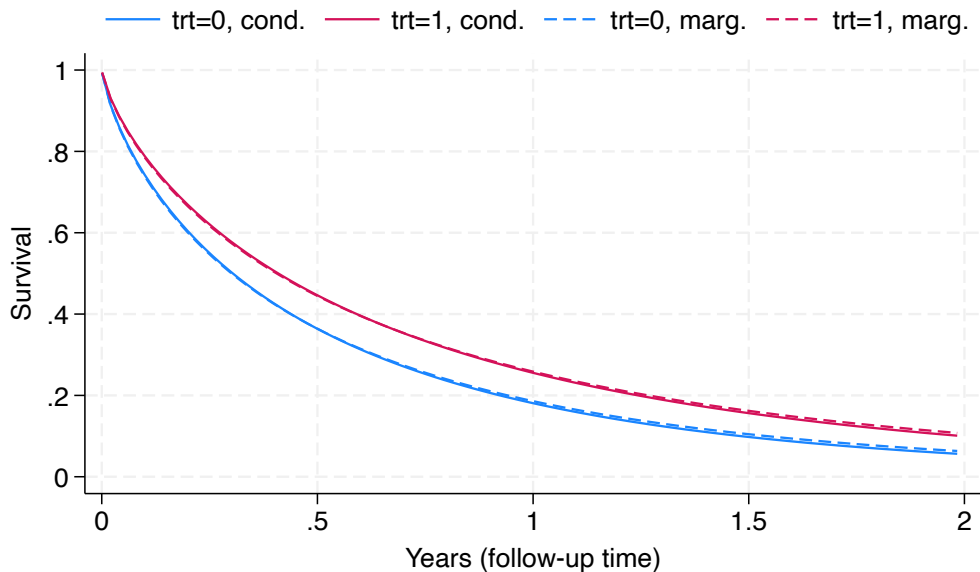

```
. stcurve, survival at1(trt=0) at2(trt=1) fixedonly ///
>          title("") legend(rows(1) position(12))
```

```
. stcurve, survival at1(trt=0) at2(trt=1) marginal ///
>         title("") legend(rows(1) position(12))
```

# Contextual Effects

Hierarchical models are useful to take into account and explicitly model the correlation between study units within the same cluster, such as each trial participants in our case study.

The higher the variance between clusters, the stronger the evidence towards the existence of a general contextual effect.

This contextual effect is named *general* because it reflects the influence of the cluster context as a whole, without specifying any contextual characteristic other than the very boundaries that delimit the cluster.

I will now introduce the following measures that are often used to quantify contextual effects:

- The estimated variance of the random effect(s);
- The best linear unbiased predictions (BLUPs) of the random effects;
- The median (across clusters) hazard ratio.

# Variance of the Random Intercept

```
. mestreg, noheader
-------------------------------------------------------------------------------
         _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
------------+------------------------------------------------------------------
[ Output omitted... ]
------------+------------------------------------------------------------------
trialID     |
  var(_cons)|  .0470106    .0234363                      .0176947    .1248956
-------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
Note: _cons estimates baseline hazard (conditional on zero random effects).
LR test vs. Weibull model: chibar2(01) = 18.97          Prob >= chibar2 = 0.0000
```

In the case study, the estimated variance of the random effects is .0470, with a 95% C.I. of: (.0177, .1249). But:

- The magnitude of the estimated variance depends on the scale of time;
- This quantity is hard to interpret as it is.

# BLUPs

A more interpretable comparison is one involving best linear unbiased predictions (BLUPs) of the random effects.

The BLUPs are predicted using a procedure that uses the fitted model coefficients and the observed data.

In loose words, predicting the BLUPs means predicting a cluster-specific effect for the latent, unobserved random effects. These predicted cluster effects can then be used elsewhere:

- In ranking,
- In further predictions,
- Etc.

# BLUPs in Stata

BLUPs (and their standard errors) can be predicted after `mestreg` using the following code:

```
. predict b_trialID, reffects reses(se_b_trialID)
(calculating posterior means of random effects)
(using 7 quadrature points)
.
. // Tag distinct random effects, useful for plotting/ranking
. bysort trialID: gen plot_tag = _n == 1
.
. // Rank the predicted random effects
. egen trialID_rank = rank(b_trialID) if plot_tag == 1
(1,142 missing values generated)
.
. // Make confidence intervals for random effect values
. gen b_trialID_lci = b_trialID - invnormal(1 - 0.05 / 2) * se_b_trialID
. gen b_trialID_uci = b_trialID + invnormal(1 - 0.05 / 2) * se_b_trialID
```

# Ranked BLUPs
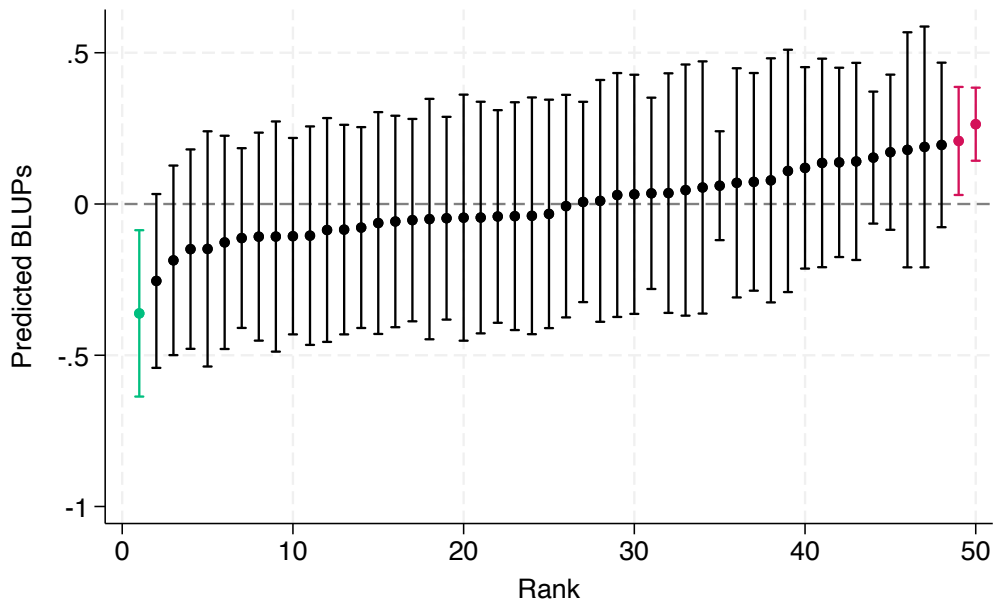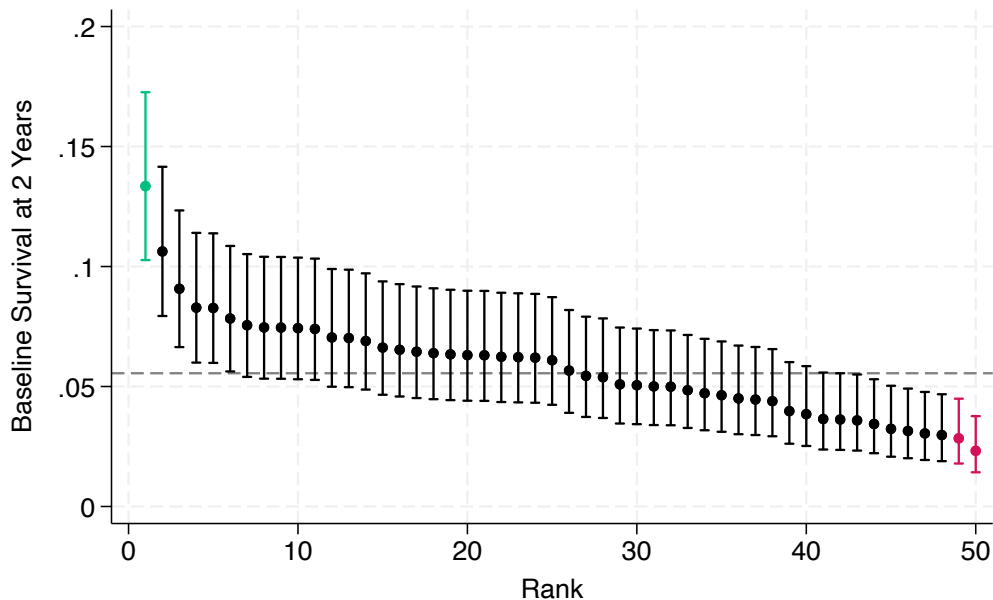
# Median Hazard Ratio

The median hazard ratio is a measure that can be for assessing the magnitude of the general contextual (i.e., trial) effects when analysing survival data.

When a random intercept follows a normal distribution, this is calculated as:

$$\text{MHR} = \exp\left[\sqrt{2\hat{\sigma}^2}\phi^{-1}(0.75)\right]$$

where $\hat{\sigma}^2$ is the estimated variance of the random intercept and $\phi^{-1}$ is the inverse of a standard normal cumulative distribution function.

Note that a MHR can take any positive value.

# Median Hazard Ratio: Estimation

The fitted variance of the random intercept can be extracted from e(b):

```
. display _b[/:var(_cons[trialID])]
.04701059
```

Then, we can calculate the median hazard ratio, and a confidence interval, using nlcom:

```
. nlcom exp(sqrt(2 * _b[/:var(_cons[trialID])]) * invnormal(0.75))

       _nl_1: exp(sqrt(2 * _b[/:var(_cons[trialID])]) * invnormal(0.75))

------------------------------------------------------------------------------
         _t |  Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
------------+-----------------------------------------------------------------
      _nl_1 |   1.229759    .0633974     19.40   0.000     1.105502    1.354015
------------------------------------------------------------------------------
```

# Median Hazard Ratio: Interpretation

```
. nlcom exp(sqrt(2 * _b[/:var(_cons[trialID]])]) * invnormal(0.75))

------------------------------------------------------------------------------
         _t |  Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
------------+-----------------------------------------------------------------
      _nl_1 |    1.229759   .0633974    19.40   0.000     1.105502    1.354015
------------------------------------------------------------------------------
```

This can be interpreted as follows:

- The median relative change in the hazard of the outcome between a subject from a trial at a higher risk for the outcome and an identical subject in a trial at a lower risk for the outcome is 1.230 (95% C.I.: 1.106, 1.354).

- This is comparable to the effect of treatment, where the hazard ratio was 0.780 (95% C.I.: 0.703, 0.905) or 1.253 (inverting treatment, 95% C.I.: 1.104, 1.422).

# Conclusions

In conclusion, today we have seen:

- How to fit hierarchical survival models using Stata's built-in `mestreg` command;

- How to obtain a variety of post-estimation predictions;

- Differences between conditional and marginal predictions;

- Examples of contextual effects quantifying differences between clusters;

Finally, the case study results can be replicated using code and data that is openly available on GitHub: https://github.com/RedDoorAnalytics/2023-nesc-hsurv