**eAppendix 1. Description of the joint model**

Repeated %FEV1 measurements and survival data were modelled using a joint model with shared random effects between the measurement sub-model and the survival sub-model (for an introduction to joint models see Asar *et al*, 2015). If denotes the distribution of the measurement process and denotes the distribution of the survival process, then a key assumption of joint models is that the measurement and survival processes are independent conditional on the random effects , i.e. the joint likelihood factorises as

(1)

The joint model was fitted by maximising the likelihood (1), thus all repeated measurement and survival parameters are estimated simultaneously and random effects are predicted by their expectations conditional on both repeated measurements and survival data.

# Model specification

Let denote the number of patients, and the number of follow-up measurements on the *i*th patient. Let denote the vector of baseline covariates (birth year, sex and coded F508 allele status), the age, and the measured value of %FEV1, for the *i*th patient at their *j*th follow-up time . Let denote the vector of interactions (, , , ). Let be independent and identically distributed bivariate Normal random effects with means zero, variances and , and correlation . Finally, let be independent and identically distributed Normal random variables with mean zero and variance.

*U*

*V*

Our measurement sub-model takes the form

(2)

where is the underlying error-free value of the th patient’s %FEV1 at age .

Our survival sub-model partitions the follow-up period for each subject into discrete, one-year intervals and assumes that the probability of survival in the th such interval is

(3)

where denotes the standard Normal distribution function, is again the vector of baseline covariates (birth year, sex and coded F508 allele status), is the age of the th patient at the mid-point of the th follow-up interval and is the rate of change in the th patient’s underlying error-free %FEV1.

Conditional on the random effects for the th patient, the probability that they deliver

measurements is

where if the patient in question dies between the scheduled times of their th and st measurements and otherwise. Also conditional on , the joint distribution of

the th patient’s measurement sequence is multivariate Normal with mean vector and variance matrix , where is the identity matrix.

*i*

The contribution of the th patient to the likelihood function for the data follows by combining

(2) with (3) and marginalising with respect to the bivariate Normal distribution of , to give

In (4), *R* is the symmetric 2 by 2 matrix with diagonal elements and , and off-diagonal

elements , and denotes the probability density of a multivariate Normal

distribution with mean vector , variance matrix and argument (Barrett *et al*, 2015). The log-likelihood is then given by the summing the logarithm of (4) over patients, hence

where *θ* denotes the complete set of model parameters.

To fit the model, we estimate by numerical maximisation of , using the R function optim with the following argument settings: method="BFGS"; reltol=1e-9; maxiter=10000. We then use the R function hessian from the numDeriv package to obtain the numerical approximation to the Hessian matrix, from which we calculate standard errors and report 95% confidence intervals as estimates *±*2 standard errors.

# Calculation of hazard ratios

To express covariate effects as hazard ratios, we approximate the hazard ratio by the risk ratio of experiencing an event in the next year

where is the estimated risk given by equation (3) using parameter estimates from the fitted model for patient 1 with covariates and , and random effects and . And similarly for .

We estimate the variance for the log HRusing the delta method

where and are the linear predictors for the probit survival model for patients 1 and 2 respectively, and

and denotes the standard Normal density function. The variance and covariance of the

linear predictors is estimated as, e.g. for

where is the variance matrix of the vector of parameter estimates , and is estimated by the inverse of the Fisher information matrix. Confidence intervals are then estimated on the log HRscale, and exponentiated to the HR scale.

# References

Asar, O¨ ., Ritchie, J., Kalra, P.A. and Diggle, P.J. (2015). Joint modelling of repeated measure- ment and time-to-event data: an introductory tutorial. *International Journal of Epidemiology*, **44**, 334–344.

Barrett, J., Diggle, P.J., Henderson, R. and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society*, B **77**, 131–148.

**eAppendix 2. Description of the joint model Exploratory plots and model checks**

**Histogram of age at entry to cohort**

**![A screenshot of a cell phone

Description automatically generated]()**

**Supplementary Figure 1:** Histogram of age at study entry.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Female** | **Male** | **All** |
| Number post-transplant (of any type) at baseline | 62 | 91 | 153 |
| Number receiving lung transplant post-baseline | 215 | 197 | 412 |
| Number of deaths amongst those receiving  lung transplant post-baseline | 68 | 47 | 115 |

**Supplementary Table 1:** Numbers of transplants by sex

**Random effect parameter estimates for joint model**

|  |  |
| --- | --- |
|  | Estimate (95% CI) |
| Random intercept S.E. | 22.24 (21.78, 22.71) |
| Random slope S.E. | 1.47 (1.44, 1.50) |
| Correlation between random intercept and random slope | -0.56 (-0.57, -0.54) |
| Residual variance in %FEV1 model | 9.23 (9.18, 9.29) |

**Supplementary Table 2:** Joint modelparameter estimates and 95% CIs for all parameters not reported in Tables 2 and 3.

**Spaghetti plots**

** A close up of a map

Description automatically generated**

**Supplementary Figure 2:** Spaghetti plots with loess smoothers for the full longitudinal data sample (left) and a random sample of 200 individuals (right)

**Residual plots**

**A close up of a map

Description automatically generatedA close up of a map

Description automatically generated**

**Supplementary Figure 3:** Residual diagnostic plot for longitudinal model - %FEV1 outcome (left) and log%FEV1 (right).

**Empirical variogram**

A screenshot of a cell phone

Description automatically generated

**Supplementary Figure 4:** Empirical variogram for observed %FEV1 measurements. The dotted line represents the estimated total variance.

A screenshot of a cell phone

Description automatically generated

**Supplementary Figure 5:** Residual diagnostic plots for survival model. For the probit model residual plots were created using the DHARMa package in R

**eAppendix 3. Testing for mediation**

The figure below shows the directed acyclic graph (DAG) for a causal mediation analysis.

**A close up of a mans face

Description automatically generated**

Baron and Kenny outline four steps in establishing mediation(1):

Step 1:  Show that the causal variable (sex) is correlated with the outcome (death).  This is demonstrated in the probit survival models (table 3), whereby female sex is associated with increased risk of death.

Step 2: Show that the causal variable is correlated with the mediator.  This is demonstrated in table 2 (the FEV1 submodel) whereby female sex is associated with an increased rate of decline of %FEV1.

Step 3:  Show that the mediator affects the outcome variable, whilst controlling for the causal variable. This is demonstrated in the probit survival models and the joint survival model in table 3 whereby worse lung function is associated with increased risk of death, whilst adjusting for sex.

Step 4:  To establish that M completely mediates the exposure-outcome relationship, the effect of exposure on outcome controlling for the mediator should be zero. This is demonstrated in our joint model where we more accurately adjust for lung function, and the lower confidence bound for the sex effect is essentially zero; but not in the probit model.

**Limitations:**

Standard approaches to using survival models for mediation by adding a mediator as a covariate to the hazard model and then seeing if the coefficient for the exposure changes is subject to some important limitations. These include the no-confounding assumptions outlined below; and also a number of modelling assumptions, because the hazards ratio is not in general a ‘collapsible’ measure. However, the latter has been shown not to be problematic when the outcome is relatively rare (13% mortality in our study).(2)

The no-confounding assumptions mean that four types of confounding are adjusted for: (1) confounding of the exposure-outcome relationship; (2) confounding of the mediator-outcome relationship; (3) confounding of the exposure-mediator association; and (4) mediator-outcome confounders also affected by the exposure.

Alternative methods for mediation analysis are continuously being developed, but have their limitations in terms of application in routinely available software, and to our knowledge these have not been routinely applied in the context of joint models.(3)

1. Mediation (David A. Kenny) [Internet]. [cited 2018 Jul 20]. Available from: http://davidakenny.net/cm/mediate.htm

2. VanderWeele TJ. Unmeasured confounding and hazard scales: sensitivity analysis for total, direct, and indirect effects. Eur J Epidemiol. 2013 Feb;28(2):113–7.

3. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol. 2013 Sep 9;dyt127.