Appendix 1

Statistical rationale for data analysis

Generalities

The data were graphically inspected and summarized appropriately, i.e. means, standard deviations, medians and interquartile ranges for continuous variables and proportions for binary and/or categorical variables. Variables' one to one transformations (such as log) have been considered when necessary –to both reduce their variability and to achieve models' appropriateness and fit. The paper targets a wide clinical audience and detailed account of raw data has been presented accordingly.

Given the retrospective *observational* nature of the study, the main concerns to address bias and confounding is based on multivariable analyses. Therefore, a strategy for developing the most parsimonious multivariable models for each outcome of interest have been adopted (this is the term used for models with the least number of predictors yet explaining the most variability in their respective outcomes).

Univariate analyses have been conducted to assess the strength of associations with the outcomes of interest and the recorded demographic, clinical and newly derived variables. Associations with resulting p-values less than 0.1 were flagged and further considered to form a multivariable/adjusted model. Multivariable analyses have been employed and, based on complete observation only, forward, backward, step-wise or a combination of them have been applied for models' selection procedure based on . Based on a similar number of complete observations, the most parsimonious model for each outcome of interest has been derived using Akaike information AIC and BIC Bayesian information selection criteria.

Models' selection criteria are no longer valid in the presence of missing data therefore the final models for each outcome were decided upon their strength of adjusted associations exhibited in the complete data analyses (analyses based on complete observations only) and/or clinical/epidemiological rationale. Patterns in the missing data have been assessed and observed data analysis (adopting the "missing data" jargon) using multiple imputation (MI) techniques adjusted to each outcome (including event data) were performed under missing at random assumption (MAR) (1-5). The distributional assumptions of the variables have been carefully checked when these analyses were conducted. The two settings, i.e. *complete* and *observed* data analyses for each outcome, exhibited little difference in the actual estimates (only their uncertainties were tightened as expected). P-values less than 0.05 are considered of statistical

significance and beside the raw presentation, the estimates were given for the final most parsimonious models for each outcome.

Time to hearing loss

Survival analysis framework was adopted to investigate the event of hearing loss in this cohort of patients (6). Time since treatment start to hearing loss has been modelled using the semiparametric Cox proportional hazard settings and a multivariable model has been derived. Right censoring was considered the end of the treatment for those who did not experience hearing loss and this has been assumed independent of the event data. Shoenfeld's residuals and visual graphs assessed the appropriateness of the proportional hazard assumption (overall and in connection with each predictor). Probability of surviving hearing loss beyond relevant time lines have been presented base on the later MI analyses (4). MI analysis have been tailored to

Creatinine and hypokalaemia

These two binary outcomes have been constructed using the raw data for creatinine and hypokalaemia which have been investigated using logistic regression techniques in lines with the general procedure presented above. Goodness of fit tests such as Hosmer-Lemeshow would be used to assess how the models fit to the complete data (p-values less than 0.05 indicate a poor fit).

The analyses have been conducted in STATA (StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP)

Statistical references

1. James Carpenter and Michael Kenward, Multiple Imputation and its Application, John Wiley 7 Sons, Ltd, 2013.
2. Ian R. White, Patrick Royston, Angela M. Wood, A simulation Multiple imputation using chained equations: Issues and guidance for practice, Statistics in Medicine, Vol 30, Issue, pg. 377-399, 20the Feb, 2011.
3. White, Ian R, and Patrick Royston. "Imputing Missing Covariate Values for the Cox Model." *Statistics in Medicine* 28.15 (2009): 1982–1998. *PMC*. Web. 12 Jan. 2017.

4. Jonathan A C Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, James R Carpenter, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, *BMJ* 2009; 338 doi: http://dx.doi.org/10.1136/bmj.b2393
5. R Little & D. B. Rubin, Statistical Analysis with missing data, Wiley Series in Probability and Statistics 2002.
6. D. R. Cox and D. Oakes, Analysis of Survival Data, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1984