Methods section

Jono Tuke

08/12/2019

# Cleaning

The majority of the cleaning was to ensure consistency of the levels within the categorical variables. All cleaning was performed in R (R Core Team 2019), mainly using the tidyverse package (Wickham et al. 2019). All code is available from www.github.com/jonotuke/TSH\_code.

# Methods

To model the difference between the different thyroid tests (FT4, TSH or FT3) with respect to their predictive power for a variety of conditions, we used a mixed effects modelling methodology. This methodology allows us to compensate for the fact that we have more than one result in the papers considered. The modelling was performed using the lme4 package (Bates et al. 2015) and the lmerTest package (Kuznetsova, Brockhoff, and Christensen 2017).

In each case, we classified each result in a paper as showing a significant result or a non-significant result. By a significant result, we mean that a given thyroid test has been shown to be associated with a given condition at a 5% significance level. We treated the result as a binary response variable with the levels success (significant) and failure (non-significant).

In each model, we accounted for the observation of multiple results within the same paper by incorporating a random intercept for each paper. The necessity of the random intercept was determined by a likelihood ratio test and also confirmed with the Bayesian Information Criterion (BIC) (Schwarz 1978).

The thyroid test (FT4, TSH or FT3)was incorporated in the model as a fixed effect predictor. This was also tested for significance with a likelihood ratio test and BIC.

For significant models, we also calculate the Tukey pairwise comparisons between the thyroid tests using the multcomp package (Hothorn, Bretz, and Westfall 2008).

We were also concerned about the potential predictive power of other covariates. We considered the clinical ``system" (eg Cardiac, Bone, Pregnancy) each condition pertained to - as classified by the physician author SF; the number of subjects in the analysis; and the number of covariates in the model of the analysis. For each of these covariates, we considered six models:

1. thyroid test,
2. thyroid test with random intercept for paper,
3. thyroid test and stated covariate,
4. thyroid test and stated covariate with random intercept for paper,
5. thyroid test and stated covariate and interaction between thyroid test and covariate, and
6. thyroid test and stated covariate and interaction between thyroid test and covariate with random intercept for paper.

The best model was chosen with BIC.

As we have further dependencies within each paper: the cohort used for the analysis; the general type of analysis; and sophistication of the models, we also tested for the necessity of a nested random intercept.

The final attempt to account for dependency we called the “Devil’s advocate” method. For each of the following strata:

1. smallest number of subjects, simple model;
2. smallest number of subjects, complex model;
3. largest number of subjects, simple model; and
4. largest number of subjects, complex model,

we randomly selected one analysis from each paper to represent the strata. We then performed a logistic regression with significance as the response variable and thyroid test as the predictor.

# Meta-analysis

The obvious question is why pursue this complicated approach - why not simply perform a traditional meta-analysis? Fundamentally, meta-analysis answers the wrong question. If we were solely interested in the predictive power of a single thyroid test with regards to a single condition, then we would use a traditional meta-analysis, but we are considering a more general question - which thyroid test is most predictive of a wide class of conditions. To recap, to perform a traditional meta-analysis we would need to compare like to like, that is we would need to select only those analyses that correspond to the same condition (eg Atrial Fibriliation or Breast Cancer), and using the same analytic methodology (eg Cox’s Proportional Hazard, or Pearson’s correlation coefficient). Given that explanation, we still did consider a traditional meta-analysis. Unfortunately, we found that once you count the number of papers considering the same condition and using the same methodology, there are very few condition-method pairs with at least two papers to compare, thus rendering meta-analysis incapable of making a contribution to answering our question.

# Results

## Overall

There are 58 papers in total.

# References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. “Simultaneous Inference in General Parametric Models.” *Biometrical Journal* 50 (3): 346–63.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software* 82 (13): 1–26. <https://doi.org/10.18637/jss.v082.i13>.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model.” *Ann. Statist.* 6 (2): 461–64. <https://doi.org/10.1214/aos/1176344136>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.