Methods section

Jono Tuke and Nigel Bean

15/12/2019

# 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, with a total of 2439 results. The number of subjects in the various analyses are summarised below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| minimum | q1 | median | mean | q3 | maximum | na |
| 18 | 1151 | 2078 | 3070.811 | 3615 | 10990 | 575 |

The observed results for the considered thyroid tests are given below.

|  |  |  |  |
| --- | --- | --- | --- |
| Thyroid\_test | Y | N | total |
| FT4 | 406 | 407 | 813 |
| T3 | 134 | 120 | 254 |
| TSH | 183 | 630 | 813 |

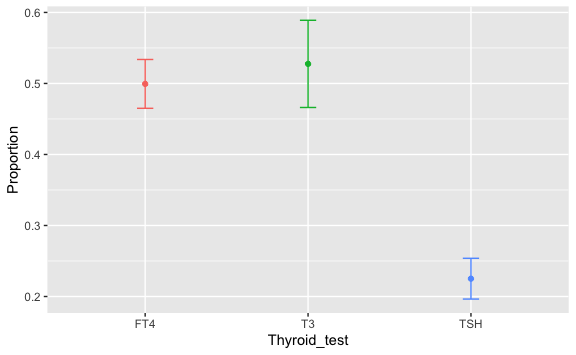
And the equivalent as proportions

|  |  |  |
| --- | --- | --- |
| Thyroid\_test | N | Y |
| FT4 | 0.50 | 0.50 |
| T3 | 0.47 | 0.53 |
| TSH | 0.77 | 0.23 |

The 95% confidence interval for the population proportions are given below (asymptotic normality confidence interval were used).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thyroid\_test | N | n\_sig | p | lwr | upr |
| FT4 | 813 | 406 | 0.50 | 0.47 | 0.53 |
| T3 | 254 | 134 | 0.53 | 0.47 | 0.59 |
| TSH | 813 | 183 | 0.23 | 0.20 | 0.25 |

These are illustrated in the following figure



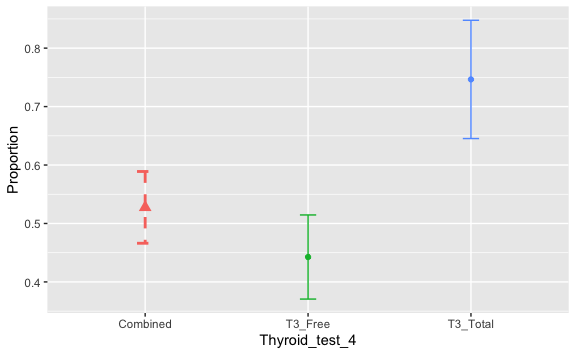
We see that both Free T4 and T3 have statistically higher significant results compared to TSH. There is no discernible difference between Free T4 and T3.

Figures of the 95% confidence interval of the proportion of significant results for each of the different thyroid tests stratified by

* system,
* condition,
* analysis methodology,
* cohourt size,
* and number of covariates in the analysis

are given in the Appendix.

We also separated T3 into free T3 and total T3. The figure is given below. We see that total T3 has a statistically higher proportion of significant results compared to free T3.



## Thyroid test as predictor

It was necessary to incorporate a random intercept for paper in the model (P-value ). There is a statistically significant effect of thyroid test in predicting a result’s significance (P-value ). Post-hoc pairwise comparisons show that there is a statistically higher proportion of significant results for free T4 compared to TSH (P- value ), and also a statistically higher proportion of significant results for T3 compared to TSH (P- value ). These results confirm those illustrated in the confidence interval plots.

## Thyroid test as predictor with T3 stratified into total T3 and free T3

Again we found a statistically significant random intercept for paper (P-value ), and a significant predictive effect of thyroid test (P-value )

Pairwise comparisons reveal the following significant results:

* Free T4 higher than TSH (P-value < 0.001),
* Free T3 higher than TSH (P-value < 0.001), and
* Total T3 higher than TSH (P-value < 0.001).

## Incorporation of other covariates

We found that the additional main effects of system, cohort size, and number of covariates did not improve the predictive effect of the model, compared to one with just thyroid test (based on minimising BIC).

## Further nested random intercepts

We found that a nested random effects structure of cohort within paper did improve the fit of the model, but did not change the observed effects of thyroid tests to that given above.

## Devils advocate

In all four Devil’s advocate model, we found a statistically significant effect of thyroid test on the proportion of statistically significant results. Pairwise comparisons reveal that the only signifcant results in all four Devil’s advocate models were Free T4 higher than TSH. The P-values are given in the following table.

|  |  |  |
| --- | --- | --- |
| Model | P-value Thryoid Test | P-value FT4 > TSH |
| smallest number of subjects, simple model | 0.0001891 | 0.000202 |
| smallest number of subjects, complex model | 0.0007730 | <0.001 |
| largest number of subjects, simple model | 0.0112600 | 0.00905 |
| largest number of subjects, complex model | 0.0006587 | 0.000638 |

## Restricted analysis

We also considered a restricted analysis where we excluded all papers that showed an incongruent result or were cross-sectional analyses. The key difference from the above observations were that the proportion of results that were significant for T3 went from 0.53 to 0.13; the full summary is in the table below. All statistical analyses in terms of the comparison between FT4 and TSH remained consistent in this restricted analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thyroid\_test | N | n\_sig | p | lwr | upr |
| FT4 | 400 | 195 | 0.49 | 0.44 | 0.54 |
| T3 | 52 | 7 | 0.13 | 0.04 | 0.23 |
| TSH | 400 | 66 | 0.16 | 0.13 | 0.20 |

# 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). The Institute of Mathematical Statistics: 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>.