Report on manual checks of randomly selected spelling errors and controls

Adrian Barnett

24 February 2025

This report shows the results of the manual checks of random samples to estimate the false positive and false negative error rates.

To check for false positives, we selected spelling errors flagged by our algorithm from *PubMed* abstracts. These randomly selected errors were confirmed as errors by manually checking the abstract in *PubMed* and the publisher’s version. A false positive is where a word labelled as an error by our algorithm is instead correct.

To check for false negatives, we randomly sampled 200 control abstracts that we hoped had no spelling errors. The abstracts were pasted into one *Word* document and checked for errors using the *Word* spell check using US spelling. We then checked each flagged error.

## Results of the manual check of false positives

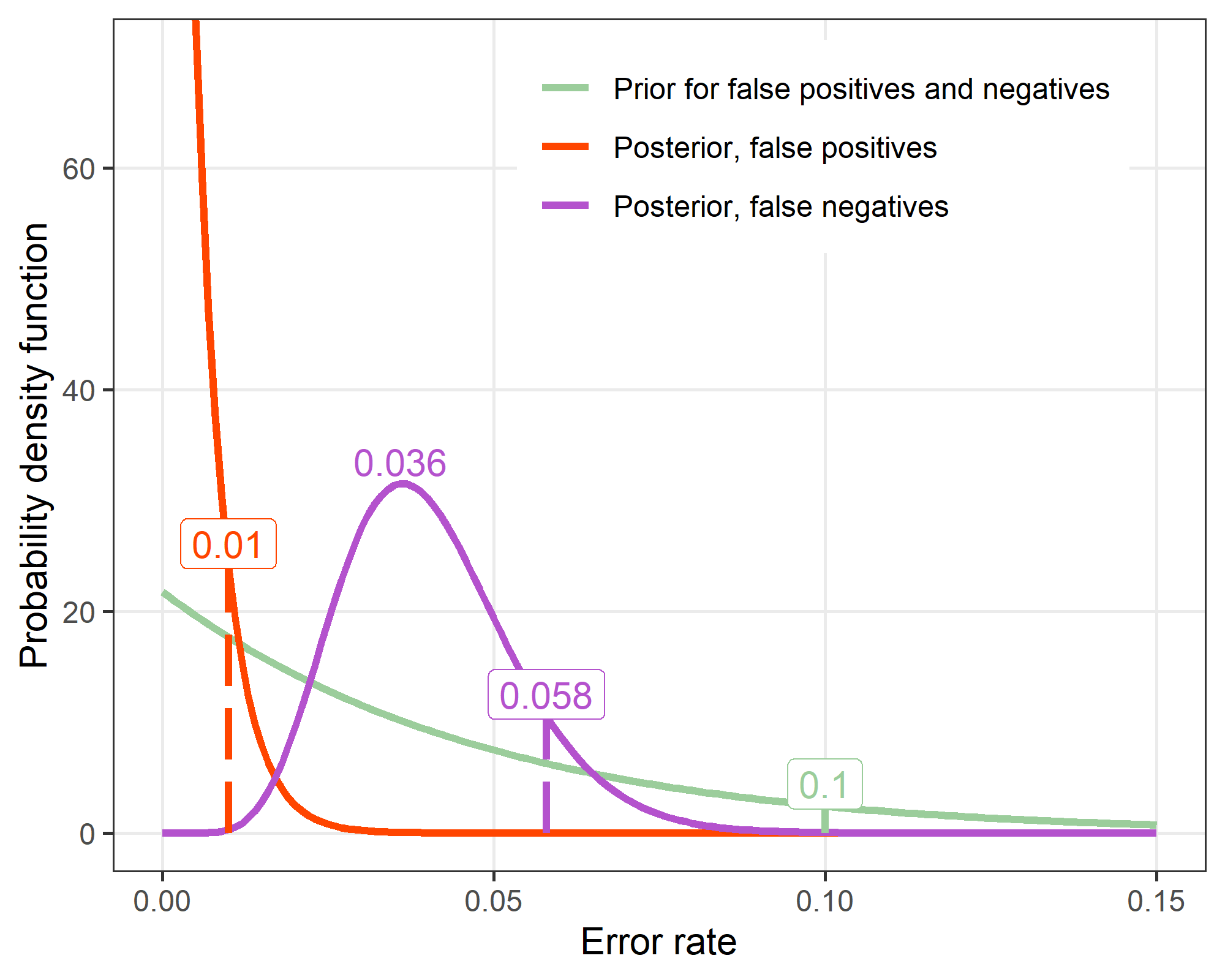
| **verified** | **n** | **Percent** |
| --- | --- | --- |
| yes | 200 | 100 |

All randomly selected spelling errors were confirmed as errors.

## Results of the manual check of false negatives

There were 8 abstracts in the 200 randomly selected controls that included a spelling error. The errors included misspelled words due to missing spaces and typos.

### Bayesian calculations for error rates



The upper limit of the y-axis is limited to 70 to avoid squashing the distributions for the prior and false negatives.

We used a Bayesian calculation using the Beta distribution to estimate the false positive and negative rates. We assumed a sloping Beta prior with a declining probability of relatively high error rates and a 90% probability that the error rate is under 0.1.

For false positives, there were 0 mistakes in the random sample of 200 abstracts. Combining these data with the prior gives the posterior distribution for false positives with a 90% probability that the false positive rate is under 0.01. The posterior mode is a false positive rate of 0.

For false negatives, there were 8 mistakes in the random sample of 200 abstracts. Combining these data with the prior gives the posterior distribution for false negatives with a 90% probability that the false positive rate is under 0.058. The posterior mode for the false negative rate is 0.036.

# Checks of variables in case–control data

To check the data on cases and controls, we randomly selected 50 cases and 50 controls from the final dataset. We checked the journal and publisher, number of authors and first author’s country, and publication date. The selections are in 7\_manual\_checks.html. There were no errors except for one publisher for PMID = 18799226 published in the journal “Trends in Biotechnology”, where the publisher in our data “Elsevier” but would have been better as “Cell Press”, although “Elsevier” is the parent company. We noted this was true for all the “Trends” journals, hence we changed the published to “Cell Press” for all these journals.