Report on manual checks of randomly selected words to verify they are spelling errors

Adrian Barnett

06 December 2024

This report shows the results of randomly selected purported spelling errors from abstracts which were then manually checked in *PubMed*. A false positive is where a word we have labelled as an error is instead correct. We did not examine false negatives.

## Results of the first check

The first check was made after just a few hundred errors had been searched for in *PubMed*.

| **verified** | **n** | **percent** |
| --- | --- | --- |
| no | 6 | 0.06 |
| yes | 94 | 0.94 |

### Reasons for mistakes in the algorithm (first check)

* was an acronym
* was an acronym
* chemical symbol
* “Ag-In alloy”
* “Ag-In alloy”
* acronym

The algorithm is not good at detecting acronyms which are sometimes based on real words with imperfect spelling, e.g., the COCTAIL study.

## Results of the second check

The second check was made after all errors had been searched for in *PubMed*.

| **verified** | **n** | **percent** |
| --- | --- | --- |
| no | 11 | 0.11 |
| yes | 89 | 0.89 |

### Reasons for mistakes in the algorithm (second check)

* acronym
* acronym
* in French
* had hyphen
* was “peri-stent”
* French
* acronym
* person’s name
* person’s name
* name “Wholey wire”
* person’s name

There were repeated problems with acronyms and people’s names, e.g. Sergent. We introduced another check of the text to avoid these errors.

## Results of the third/fourth check

The third and fourth checks were made after the additional step to remove acronyms and names.

| **verified** | **n** | **percent** | **valid\_percent** |
| --- | --- | --- | --- |
| no | 5 | 0.01666667 | 0.01945525 |
| yes | 252 | 0.84000000 | 0.98054475 |
|  | 43 | 0.14333333 |  |

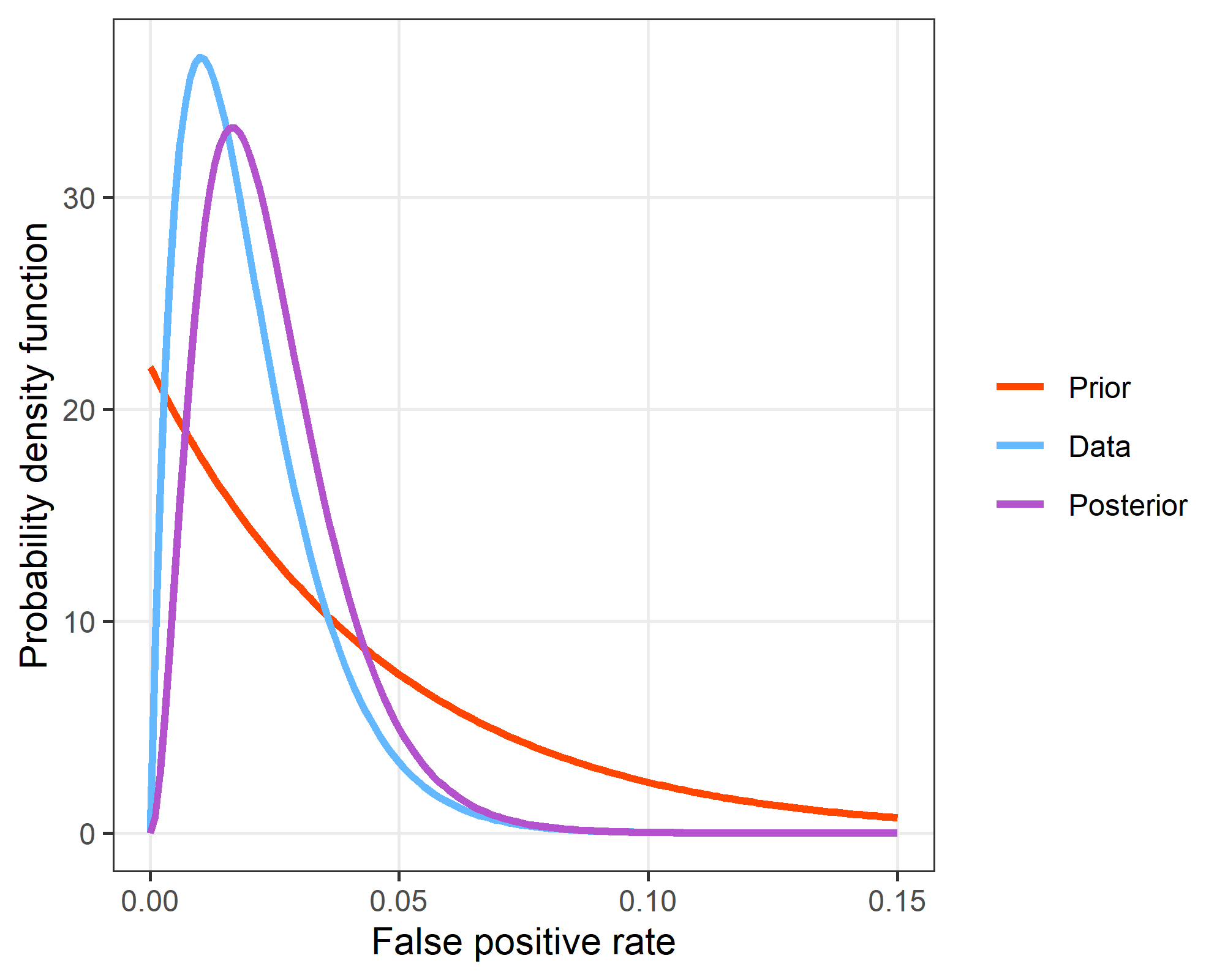
### Reasons for mistakes in the algorithm (third and fourth checks)

* is a word
* is a word: “sheat-fish”

There were no acronyms or names in the third/fourth check.

The two non-errors were excluded from the data.

### Bayesian calculation



We used a Bayesian calculation using the Beta distribution to estimate the upper false positive rate. 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. The Beta distribution for the data used 2 failures and 98 successes. For the posterior, there is a 90% probability that the false positive rate is under 0.043. The posterior mode is a false positive rate of 0.017.