# Appendix A. Calculating post-test probabilities with known sensitivity, specificity and base rate

Bayesian hypothesis testing relies on Bayes’ theorem, which helps solve problems such as the following:

“A patient is referred to a memory clinic to evaluate whether he has an objective memory impairment. Of the 40 patients referred to the memory clinic, 4 have an objective memory impairment of whom 2 have a positive test result (i.e., a test result that exceeds a pre-specified criterion). Out of the 36 patients referred to the memory clinic without an objective memory impairment, 4 patients have a positive test result. If a patient has a positive test result, what is the probability that he has an objective memory impairment?”

One may be inclined to state that the patient has a 50% probability to have an objective memory impairment, given that 2 out of 4 patients with an objective memory impairment have a positive test result. However, the correct answer is actually 33%. People’s intuitions about these problems indeed tend to deviate from the true result (Binder et al., 2015). It often helps to solve these problems using a two-by-two table or visualization such as in Figure A1 (Binder et al., 2015). To obtain the correct result for such problems, one needs to apply Bayes’ theorem:

Equation A1. .

The *posterior probability* () refers to the probability that a certain hypothesis (H) is true given the observed data (D). In diagnosis this could reflect the probability that a patient has a COI (COI +) given a positive test result (Test +; a test result that exceeds a pre-specified criterion) or the probability that a patient does not have a COI (COI -) given a negative test result (Test -; a test result that does not exceed a pre-specified criterion). These posterior probabilities have been coined the *post-test probabilities* (Crawford et al., 2009).

Let’s first consider the situation in which the patient obtains a positive test result. We are interested in the probability that this patient has a certain COI given the positive test result (p(COI+|Test+)). Bayes’ theorem clarifies that this posterior probability can be obtained by combining the *prior probability* of the hypothesis (p(H)) with the *likelihood* (p(D|H)). In this situation, the prior probability of the hypothesis refers to the belief that the patient has a certain COI *prior* to neuropsychological testing. The best way to estimate this probability would be to base it on the prevalence of the COI (p(COI+)), also known as the *base rate*. In the example above, 4 out of 40 patients referred to a memory clinic have an objective memory impairment, constituting a 10% prior (pre-test) probability of the COI (Figure A1).

This prior probability must then be combined with the likelihood. The likelihood (p(D|H)) refers to the probability of a positive test result given the presence of the COI (p(Test+|COI+)). This likelihood is better known in neuropsychology as the *sensitivity* of the neuropsychological test (Elwood, 1993). In the example stated above, 2 out of 4 patients with an objective memory impairment have a positive test result, constituting a sensitivity of 50% (Figure A1). Last, we need to estimate p(D), representing the probability of a positive test result (p(Test+)). This probability can be obtained by combining the expected number of positive tests among the patients with the COI and the expected number of positive tests among the patients without the COI. In the problem stated above this equals 15% (Figure A1). When we combine these numbers according to Equation 1, we obtain a posterior probability of 33% (Figure A1).

We can also infer the posterior probability that a patient does not have a COI given a negative test result (p(COI-|Test-)). In this situation, the prior probability refers to our prior belief that the patient does not have a COI, which is reflected by 1 *minus* the base rate. In the problem stated above, this probability equals 90%. The likelihood refers to the probability of a negative test result given the absence of the COI (p(Test-|COI-)), which is better known as the *specificity* of the test instrument (Elwood, 1993). The specificity in the problem above equals 89%, as 32 of the 36 patients without an objective memory impairment have a negative test result (Figure A1). Last, p(D) refers to the probability of a negative test result (p(Test-)), which equals 85% in the problem above, as 2 out of 4 patients with an objective memory impairment obtain a negative test result and 32 out of 36 patients without an objective memory impairment obtain a negative test result. Thus, the posterior probability that a patient does not have an objective memory impairment given a negative test result equals 94%.

Note that in this example we treat the base rate, sensitivity and specificity as known quantities. In reality, the uncertainty of these three quantities must be considered. Methods to consider the uncertainties of these quantities have been developed by Crawford et al. (2009). In addition, we assume a perfect gold standard as reference. In reality, one must consider that the accuracy of the gold standard assessment will impact the accuracy of the estimates of the sensitivity, specificity and base rate and consequently impact the accuracy of the estimates of the post-test probabilities.

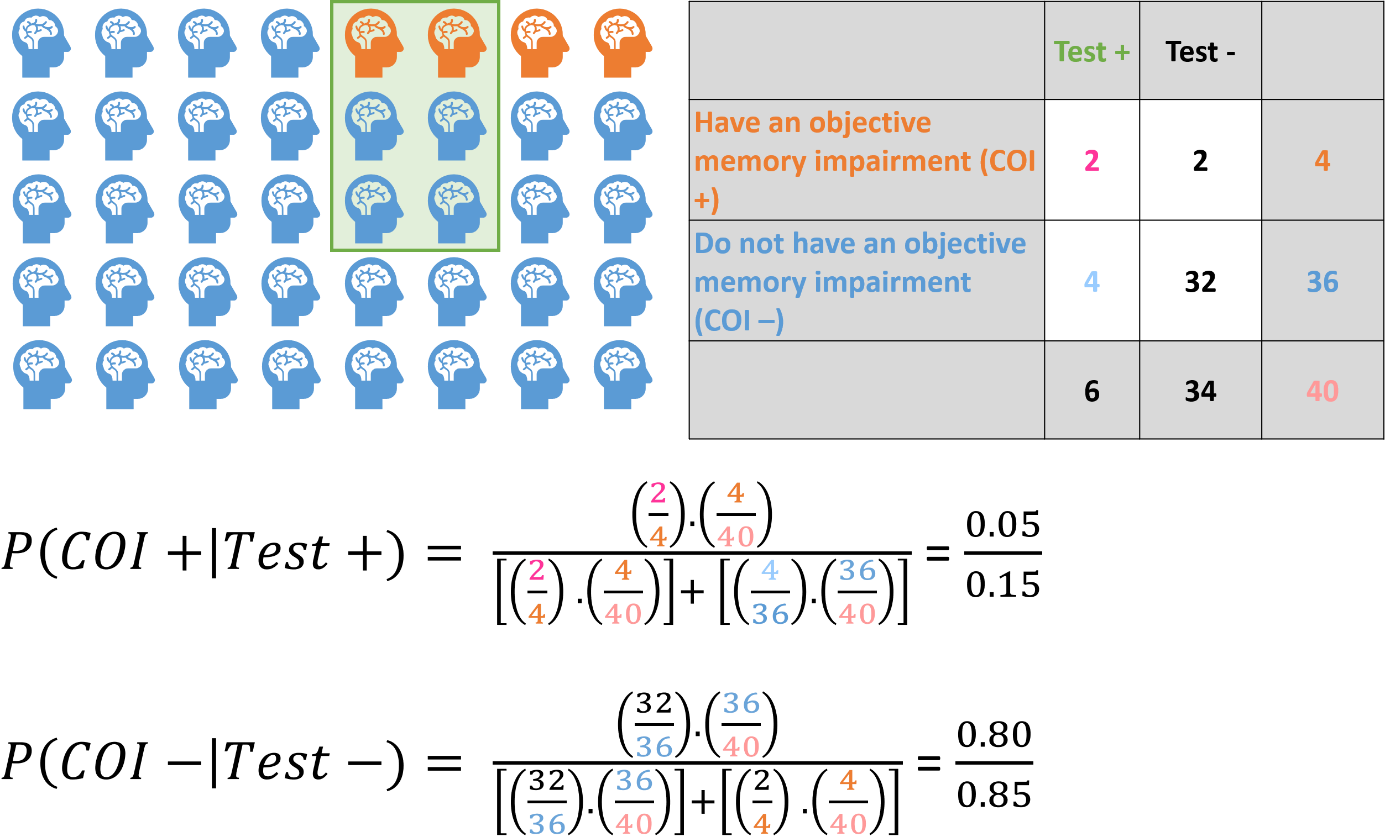


Figure A1. An illustration and 2x2 table of the problem stated above and the formula used to obtain the correct result. COI = condition of interest.

Afbeelding met schermafbeelding, monitor, telefoon, mobiele telefoon

Automatisch gegenereerde beschrijving

Figure A2. Overview of the main formulas and their relation to the 2x2 table.

# Appendix B. Calculating a Bayes Factor with known normal distributions

In the traditional diagnostic approach, neuropsychologists evaluate the probability of observing a test score equal or more extreme than a person’s test score relative to a normative group of which it is assumed that they do not have a certain condition of interest (H0). However, this approach does not inform us on the extent to which a person’s test score is likely to occur in a patient group. We will illustrate this problem in a hypothetical scenario where we use a test for which we know the distribution of test scores for the normative (H0) and the patient group (H1). Note that this scenario does not occur in reality, but it can still provide valuable insight in the limitations of z-scores as a means to inform diagnosis.

### Calculation

Consider that you administer a test of which we know the population mean and population standard deviation in the normative and in the patient group. Consider now that you aim to evaluate whether a person belongs to a patient group or a normative group. We can reformulate this question as follows:

1. A person belongs to the normative group: ,
2. A person belongs to the patient group: ,

where represents the person’s test score, represents the normal distribution, represents the population mean and represents the population standard deviation. The subscript *n* refers to the normative group, and subscript *p* refers to the patient group.

To evaluate the strength of evidence in favour of these two hypotheses, we can calculate the BF10 using the probability density functions of the normal distributions:

Equation B1.

*with*

Equation B2.

*and with*

Equation B3.

### Numerical Examples

Consider that you administer an intelligence test with known population mean of 100 and a known population standard deviation of 15 in the normative group. You administer this test to a person who scores 70. A score of 70 corresponds to a z-score of -2. The latter indicates that this score is unlikely to occur if the person would belong to the normative group.

However, you aim to know whether the person’s test score indicates that the person either belongs to a specific patient group or the normative group. Consider, a first patient group with a mean of 93 and standard deviation of 15 (Figure B1.A, left panel). To evaluate to which extent the person’s test score best matches either one of the two groups, we calculate the BF10 using the probability density functions of the normal distributions. For this scenario, the BF10 equals 2.28. Thus, based on the person’s test score it is unclear whether the person belongs either to the patient or normative group.

In contrast, the second patient group has a population mean of 78 and standard deviation of 15 (Figure B1.A, right panel). In this second scenario, the BF10 equals 6.41. The latter indicates that the person has 6.41 times greater odds as before the test that the person belongs to the patient group. Thus, although the z-score is exactly the same (-2) in both scenarios, the strength of evidence in favour of the person belonging to the patient group varies considerably.

We have now considered an example where the distributions of the patient and normative group have equal standard deviations. The relation of the z-score and BF becomes more complicated when the variances of both groups differ (Figure B1.B and Figure B1.C). That is, when the standard deviations differ, the BF does not have a monotonous relation to the z-score (Figure B2). In these scenarios it is possible that a low z-score corresponds to a BF10 that is more indicative of the normative than patient group (Figure B1.B, left panel) and that a high z-score corresponds to a BF10 that is more indicative of the patient than normative group (Figure B1.C, left panel). The latter reveals the limitations of using z-scores to inform diagnosis.

Note that this method can be further extended for scenarios where the population parameters are not known and for non-normal distributions, but this lies beyond the scope of the current illustration. Scripts for our simulations are available on https://doi.org/10.6084/m9.figshare.14500014.

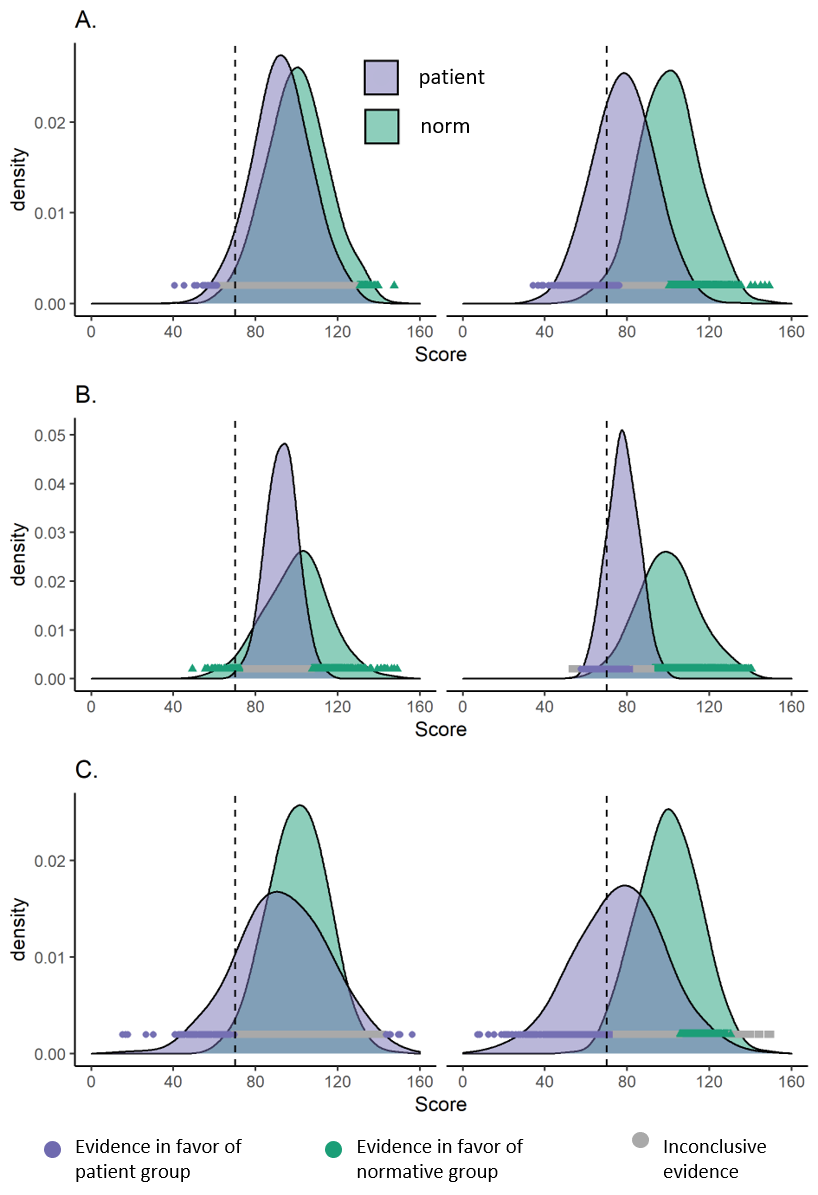


Figure B1. Distributions of test scores for the patient group (purple) and normative group (green). The dashed line represents the test score of 70, corresponding to a z-score of -2. The colored dots represent the BF for each test score. A BF10 > 3 = purple dot, BF10 < 0.33 = green triangle and BF10 in between 0.33 and 3 = grey square. Panel A) situation in which the population standard deviation of the patient and normative group is the same, panel B) situation in which the population standard deviation of the patient group is smaller than the normative group, panel C) the population standard deviation of the patient group is larger than of the normative group. Left panel represents high overlap between patient and clinical group (population mean of patient group is 0.5 SDs smaller than population mean of normative group) and right panel represents less overlap between both groups (population mean of patient group is 1.5 SDs smaller than population mean of patient group).

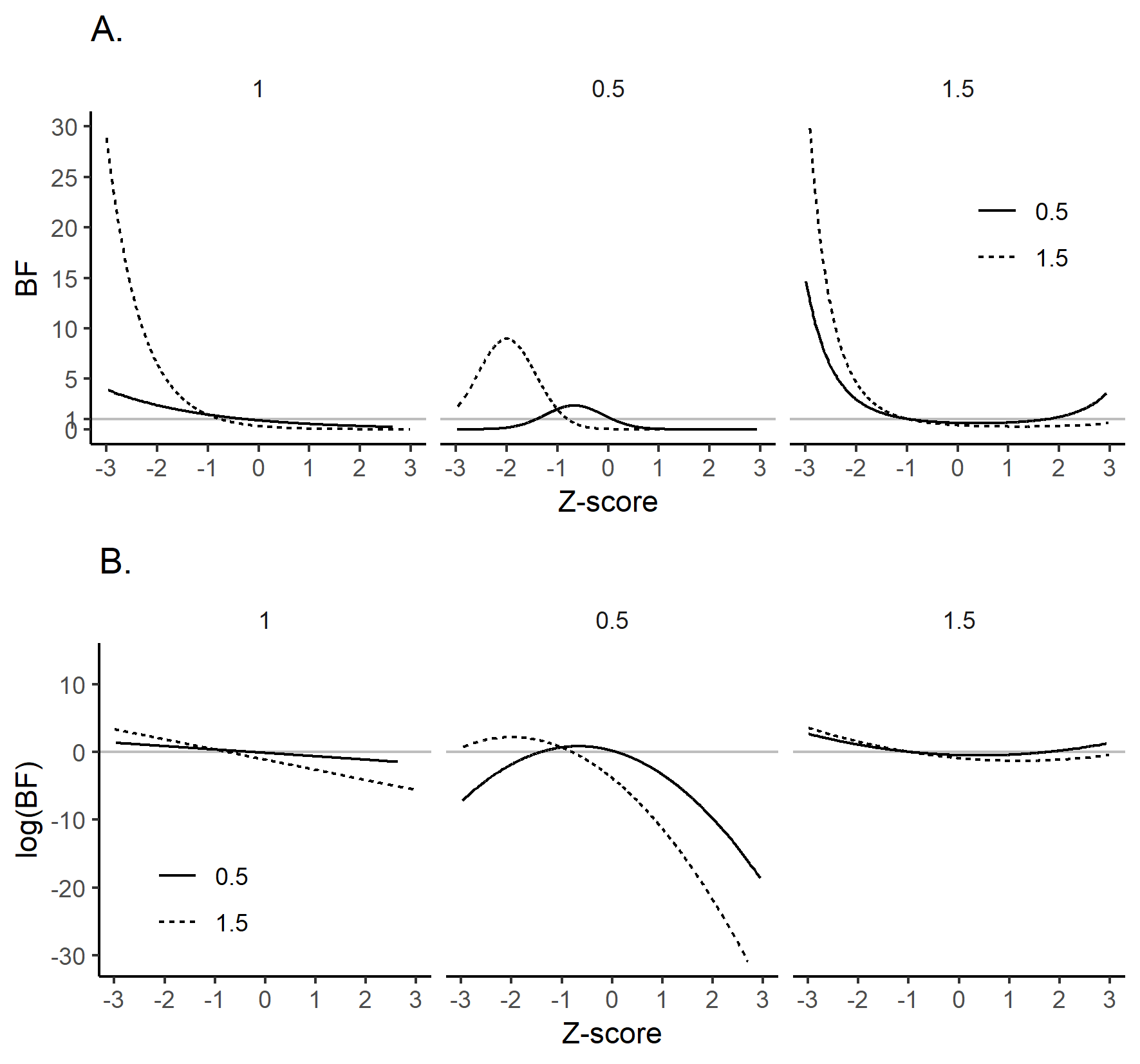


Figure B2. Relation between the BF and z-score (A) and log(BF) and z-score (B). Three scenarios are depicted: the patient and normative group have equal population standard deviations (left panel), the patient group has a smaller population standard deviation (middle panel) and larger population standard deviation (right panel) than the normative group. The solid line represents a small difference (population mean of patient group is 0.5 SDs smaller than the population mean of the normative group) and dashed line a big difference (population mean of patient group is 1.5 SDs smaller than the population mean of the normative group). The solid grey line indicates a BF of 1 (A) and log(BF) of 0 (B).

# Appendix C. Bayes Factor calculation for the cancellation test

To diagnose hemispatial neglect after stroke, cancellation tests are a popular instrument. In a cancellation test a page with shapes is presented to a patient. The patient is typically asked to mark certain shapes (i.e., targets) and ignore other shapes (i.e., distractors). The target shapes are uniformly distributed across the page. To quantify hemispatial neglect, performance on a cancellation test is often summarized by calculating the difference between the number of cancelled targets on the left and right side of the page (R-L score). The observed R-L score of the patient is then compared to percentile cut-offs that are based on the R-L score distribution in neurologically healthy individuals. However, previous research indicated that this procedure ignores binomial error variance and that it is potentially more accurate to adjust percentile cut-offs for the total performance on the cancellation test or to use a statistical test of proportions (Huygelier et al., 2020).

### Numerical Examples

In the following example we assume a known base rate probability of hemispatial neglect of 50%. Consider that you administer a cancellation test with 20 targets on each side of the page. You test two patients. The first patient cancels 18 targets on the right side and 8 targets on the left side. Thus, the patient has a R-L score of 10, which is above the normative percentile cut-off of 2 (e.g., Demeyere et al., 2015; Huygelier et al., 2019), indicating hemispatial neglect. For this patient the BF10 equals 98, indicating that it is 98 times more likely that performance differed between the left and right side than that it did not differ. Thus, in this scenario we can be confident that the patient has hemispatial neglect based on the cancellation test.

Now consider a second patient who cancels 18 targets on the right side and 16 targets on the left side. The R-L score is 2, which does not exceed the normative cut-off and it would therefore be tempting to conclude that the patient does not have hemispatial neglect. However, the BF10 for this scenario equals 0.70, indicating inconclusive evidence and a need for further assessment. Consider that we follow-up with a cancellation test with 200 targets on each side of the page. In this follow-up assessment the patient cancels 180 targets on the right side and 160 on the left side. In this follow-up assessment, the BF10 indicates that it is 8.8 times more likely that performance differed between the left and right side than that it did not differ. Thus, now it has become more clear that the patient has hemispatial neglect.

### Bayes Factor calculation

To calculate the Bayes Factors for these examples one first needs to organize the data of an individual patient on the cancellation test as a 2x2 table (Figure C1).

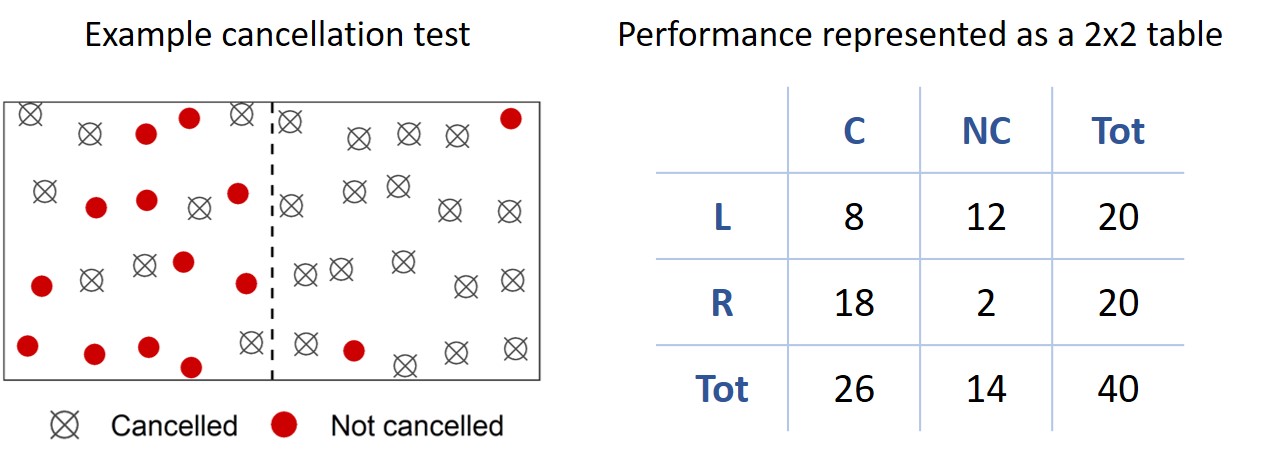


Figure C1. Example of performance of one patient on a cancellation test with 40 targets and the corresponding 2x2 table. L = left, R = right, C = cancelled, NC = not cancelled.

Given that the observed cancelled targets are likely binomially distributed, we need a test to compare two proportions (proportion cancelled targets on the left side and proportion cancelled targets on the right side). To compare these two proportions we need a Bayesian contingency table test (Jamil et al., 2017). A default Bayes Factor calculation for this test has been developed by Gunel and Dickey (Gunel & Dickey, 1974) and implemented in the BayesFactor R package and JASP software (Jamil et al., 2017). The Bayes Factor for a 2x2 table can be calculated manually using the formula (Figure C2) derived from Jamil et al. (2017). Given that there is a fixed number of targets for the left and right side of the test, we must use a test with an independent multinomial sampling plan with fixed row margins (Jamil et al., 2017). The prior distribution is a Dirichlet function in which all parameter values are equally likely (Jamil et al., 2017). In the example of the cancellation test, y1. is the total number of left targets and y2. is the total number of right targets (Figure C2). Y.1 is the number of cancelled targets and Y.2 is the number of omitted targets. Based on these row-, column and cell counts, one can calculate the BF.

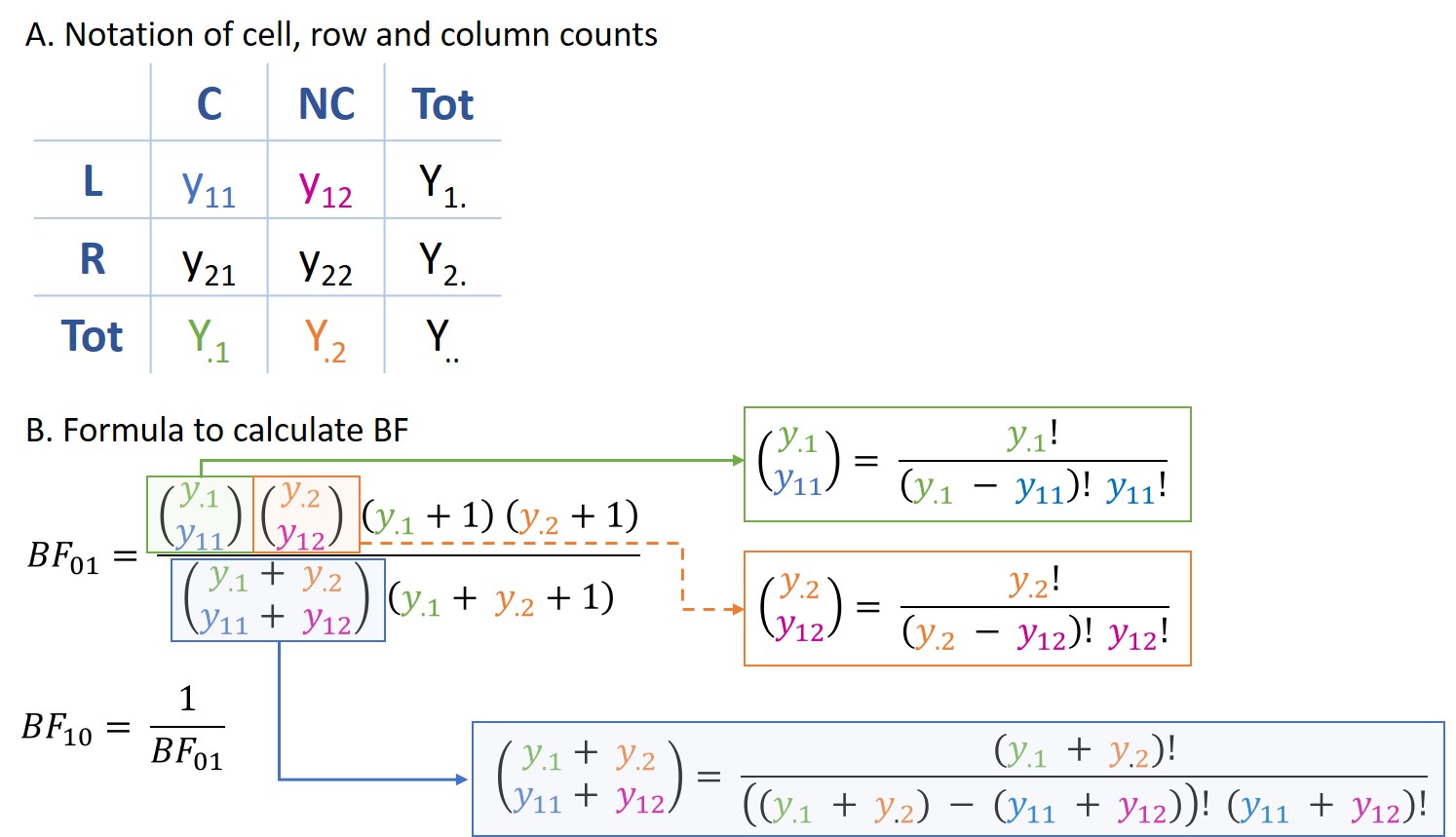


Figure C2. Formula to compute the Bayes Factor for a 2x2 contingency table with a fixed number of observations in the rows. In panel A, the notation of cell, row and column counts is visualized in relation to the 2x2 table. In panel B, the formula to calculate the BF is shown. BF01 = strength of evidence in favor of null hypothesis (performance does not depend on hemifield), BF10 = strength of evidence in favor of alternative hypothesis (performance depends on hemifield). L = left, R = right, C = cancelled, NC = not cancelled.

# Appendix D. Conditional testing

We simulated data for two hypothetical cases. *Case 1* is a patient without a spatial bias (H0) who has a 50% probability to cancel targets on both sides of the display. *Case 2* is a patient with a spatial bias (H1). The size of the spatial bias for Case 2 varied from 10% to 90% in steps of 10% difference in the probability to detect targets on the left versus right side. We simulated performance for two cancellation tests that differed in the number of targets. *Cancellation Test 1* (screen) consists of 50 targets in total (25 on the left, 25 on the right). *Cancellation Test 2* (further assessment) consists of 300 targets in total (150 on the left, 150 on the right). In this simulation we assume a known base rate probability of hemispatial neglect of 50%. Cancellation data was simulated for each scenario (i.e., each combination of a case and cancellation test), 5000 times. Cancelled hits for the left and right side were drawn from a binomial distribution (Huygelier et al., 2020).

To evaluate the presence of a spatial bias on the two cancellation tests we used a frequentist test of proportions and Bayesian contingency table test (Gunel & Dickey, 1974). If the frequentist test of proportions was statistically significant (p-value < .05), neglect was diagnosed and if it was not statistically significant, neglect was not diagnosed. Note that this reflects an erroneous statistical reasoning (accepting H0 when p-value > 0.05) but resembles current diagnostic reasoning based on normative cut-offs. If the BF10 was larger than 3, patients were diagnosed with neglect and smaller than 1/3 patients were not diagnosed with neglect. If the BF10 lied between 1/3 and 3, the results were interpreted as *inconclusive*.

In *scenario 1*, the patients who had a significant spatial bias according to the frequentist chi square test on Cancellation Test 1 (screening) were further assessed with Cancellation Test 2 to confirm the diagnosis. In *scenario 2*, the patients who did not have a significant spatial bias according to the frequentist chi square test on Cancellation Test 1 (screening) were further assessed with Cancellation Test 2 to confirm the diagnosis. The result of Cancellation Test 2 was used as the final diagnosis if it was administered. In the *Bayesian screening scenario*, the patients who had an inconclusive BF on Cancellation Test 1 (screening) were further assessed with Cancellation Test 2 to further evaluate the diagnosis. Patients with a conclusive result on Cancellation Test 1 were not further assessed.

Results of this simulation are reported in the main text (Figure 5). Scripts for our simulations are available on (https://doi.org/10.6084/m9.figshare.14500014).

# Appendix E. Sequential testing

We simulated data for several hypothetical cases. *No Neglect cases* are patients without a spatial bias (H0) who have a 50% probability to cancel targets on both sides of the display. *Neglect cases* are patients with a spatial bias (H1). The size of the spatial bias was either equal to 10%, 20%, 50% or 90%. Cancellation data was simulated for each of these hypothetical cases for different testing scenarios, 5000 times. Cancelled hits for the left and right side were drawn from a binomial distribution (Huygelier et al., 2020).

For each case, a maximum of 12 cancellation tests, each consisting of 50 targets in total (25 left, 25 right) were administered. After each cancellation test, data of all the administered cancellation tests was combined. To evaluate the presence of a spatial bias on the two cancellation tests we used a frequentist test of proportions and Bayesian contingency table test (Gunel & Dickey, 1974). If the frequentist test of proportions was statistically significant, neglect was diagnosed (i.e., reject H0). Thresholds to determine statistical significance were determined on an a priori power analysis assuming 12 tests, controlling false positives to a maximum of 5% and without futility testing (early stopping if study is unlikely to produce significant results) using the GSDesign package in R (Anderson, 2020). The following thresholds for the z-score were determined for the subsequent interim test: 3.38 (first interim test), 3.25, 3.11, 2.97, 2.83, 2.69, 2.54, 2.40, 2.24, 2.08, 1.92, 1.75 (last interim test). If the BF10 was larger than 6, patients were diagnosed with neglect and smaller than 1/6 patients were not diagnosed with neglect. If the BF10 lied between 1/6 and 6, the results were interpreted as *inconclusive*. The BF thresholds for sequential testing were based on Schönbrodt and Wagenmakers (Schönbrodt & Wagenmakers, 2018).

In addition, the sequential testing scenario was simulated under two conditions. A first condition is one in which all patients complete all planned tests. A second condition is one in which patients do not necessarily complete all planned tests (i.e., early termination). To simulate early termination, we assumed that the probability to stop data collection would increase exponentially as the number of tests increased (Figure E1). Thus, we assumed that 9% of patients would stop participation after the first cancellation test, while 67% of patients would stop participation after the 12th cancellation task.

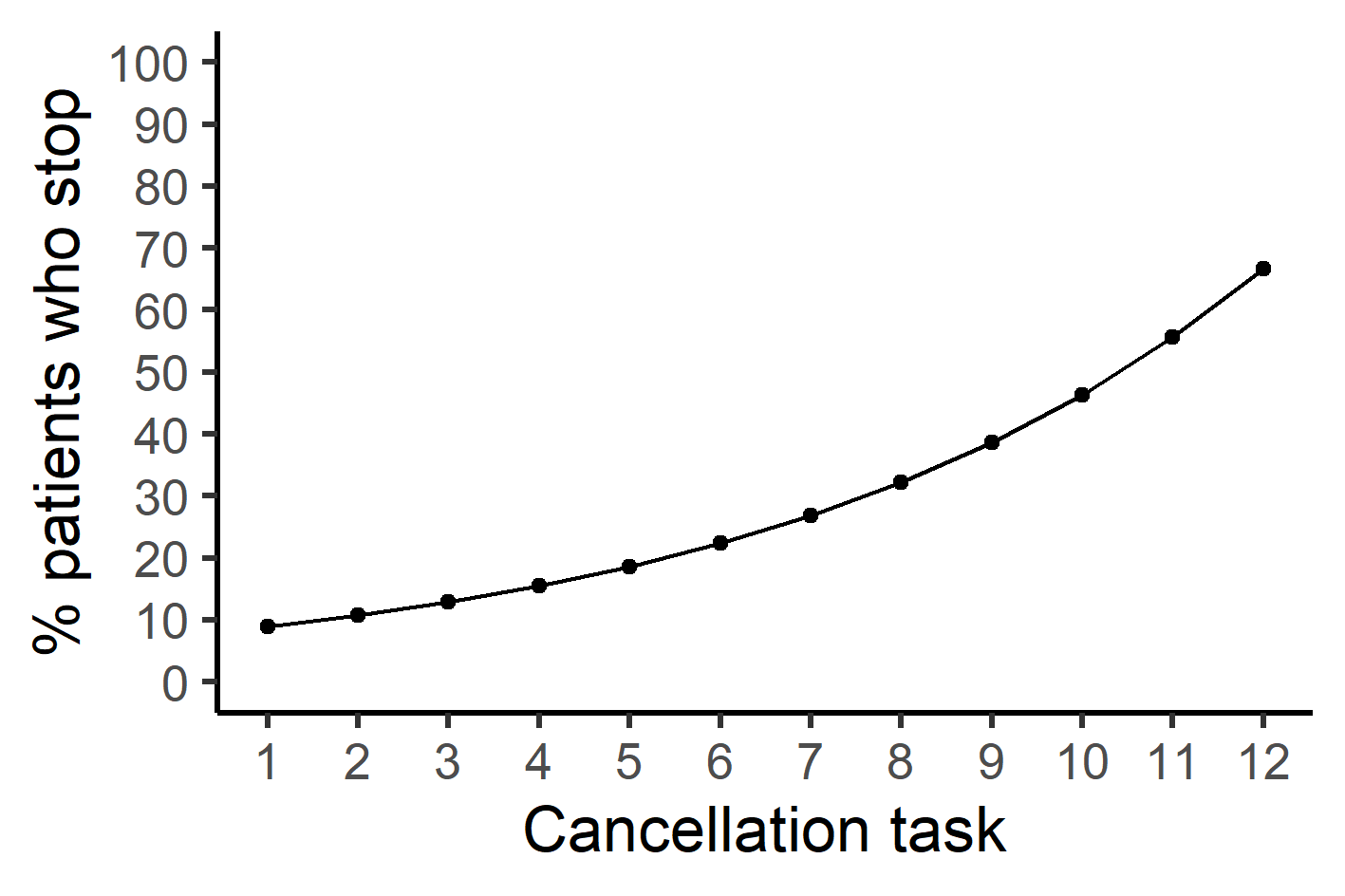


Figure E1. Percentage of patients who stop after each cancellation task.

Results of this simulation are reported in the main text (Figure 6). Scripts for our simulations are available on (https://doi.org/10.6084/m9.figshare.14500014).