Research that misleads

Henry Olders, P. Eng., MD, FRCPC
Assistant Professor, Department of Psychiatry,
McGill University
Staff Psychiatrist, Ste. Anne's Hospital

The problem

- articles with research protocols which bias results
- articles whose abstract does not reflect the entire article
- articles with conclusions not supported by the findings

Problem research protocols

- Example: Bullock R, Touchon J, Bergman H et al. Rivastigmine and donepezil treatment in moderate to moderately-severe Alzheimer's disease over a 2-year period. Curr med res opin 2005 21(8):1317-27

I want to focus on study published last year, which I believe illustrates the last problem, that of a research protocol which biases the results.

Outline of research protocol

- double-blind, randomised, controlled, multicentre trial
- evaluate the efficacy and tolerability of cholinesterase inhibitor treatment in patients with Alzheimer's disease
- 2-year period

study design

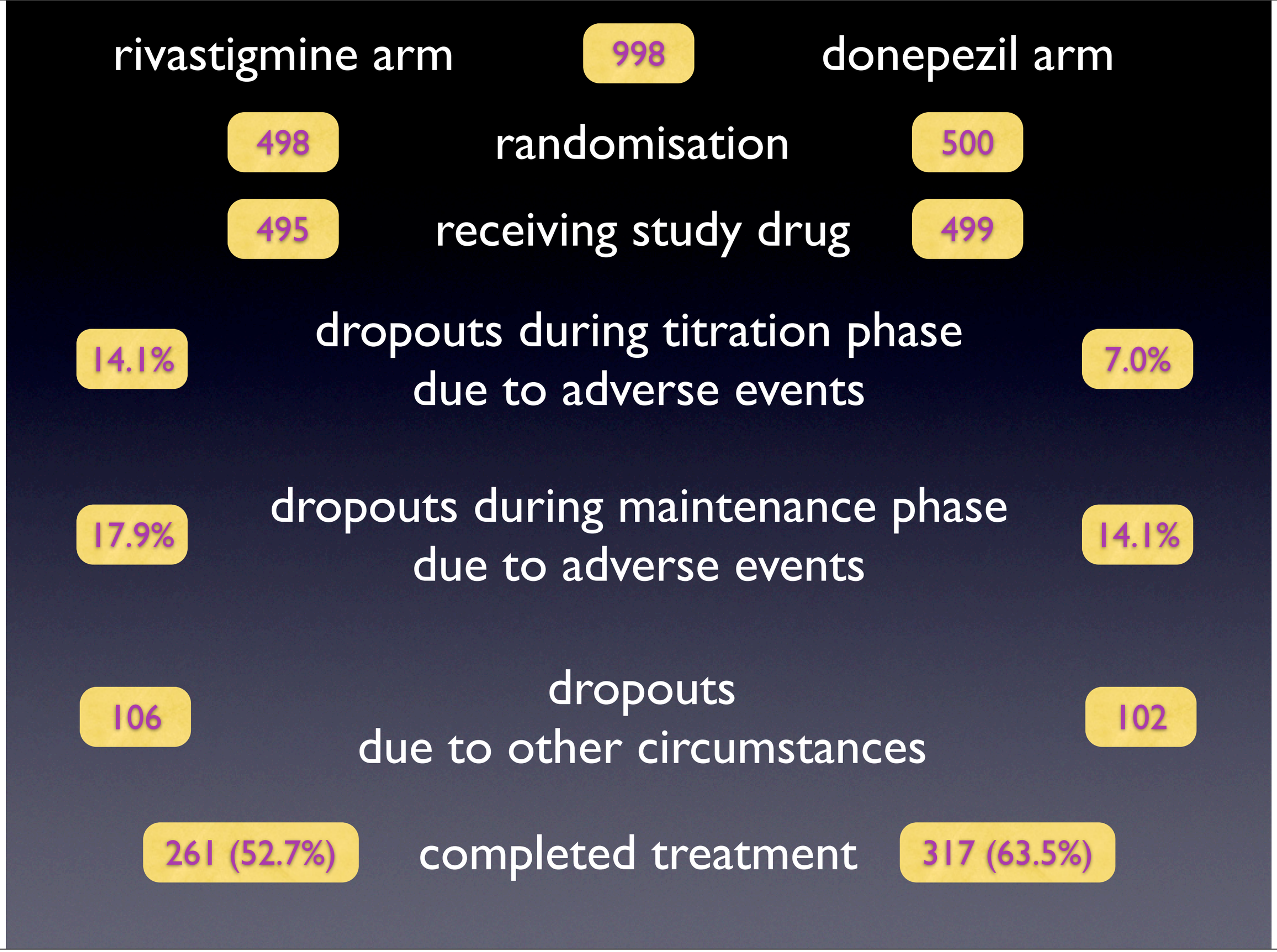
- patients randomised to treatment with either donepezil or rivastigmine
- randomised into two strata: those with MMSE 10-14, those with MMSE 15-20
- 16-week titration period, doses titrated at intervals of 4 weeks; 4 dose levels
- primary efficacy measure: Severe Impairment Battery (SIB)

study results

- 994 patients treated; 57.9% completed study
- “rivastigmine & donepezil had similar effects on measures of cognition and behaviour”
- “rivastigmine showed a statistically significant advantage on measures of activities of daily living and global functioning in the ITT-LOCF population”

conclusions

- “Cholinesterase inhibitor treatment may offer continued therapeutic benefit for up to 2 years in patients with moderate AD. Although both drugs performed similarly on cognition and behaviour, rivastigmine may provide greater benefit in activities of daily living and global functioning.”



I will go into a bit of detail into how the study was done.

Here's the patient flow. They began with 998 patients who met inclusion criteria. They were randomised either to the rivastigmine arm, or to the donepezil arm. Almost all of these patients were begun on the study drug. As was to be expected, there were people who dropped out of the study for various reasons. The article provides a breakdown of the number of dropouts for a whole list of possible reasons, but I will focus here only on dropouts due to side effects of the medication, broken down by those who dropped out during the initial 16-week dose titration phase, and dropouts during the subsequent maintenance phase lasting 88 weeks, and a third group lumping all the dropouts due to other reasons, in either phase of the study. Why will become apparent later.

14.1% of the 495 patients receiving rivastigmine dropped out due to side effects during the initial titration phase, while only 7% of the donepezil group dropped out due to side effects. This is less than half the dropout rate. During the maintenance phase, there were 17.9% side effect dropouts in the rivastigmine arm, compared to 14.1% in the donepezil arm. Keep in mind that the denominator to calculate the dropout rate at this point does not include the people who already dropped out during the titration phase, otherwise the difference between the two arms would be more pronounced.

The number patients who dropped out due to causes other than adverse effects were almost the same in the two arms, 106 and 102. That leaves 261 patients, or 52.7% of those who started on rivastigmine, completing treatment in the rivastigmine arm, and 317, or 63.5% of those started on donepezil, that completed the full 104 weeks of treatment.

Although the paper does not give the actual numbers for the patients who dropped out due to adverse effects during the titration phase or the maintenance phase, these are easy to calculate (click on the 14.1%).

rivastigmine arm

998

donepezil arm

498

randomisation

500

495

receiving study drug

499

495 x 14.1% = 70

dropouts during titration phase
due to adverse events

7%

17.9%

dropouts during maintenance phase
due to adverse events

14.1%

106

dropouts
due to other circumstances

102

261 (52.7%)

completed treatment

317 (63.5%)

In the rivastigmine arm, 70 patients dropped out due to adverse effects during the titration phase. Contrast this with the donepezil arm (click on 7% button)

rivastigmine arm

998

donepezil arm

498

randomisation

500

495

receiving study drug

499

495 x 14.1% = 70

dropouts during titration phase
due to adverse events

499 x 7.0% = 35

17.9%

dropouts during maintenance phase
due to adverse events

14.1%

106

dropouts
due to other circumstances

102

261 (52.7%)

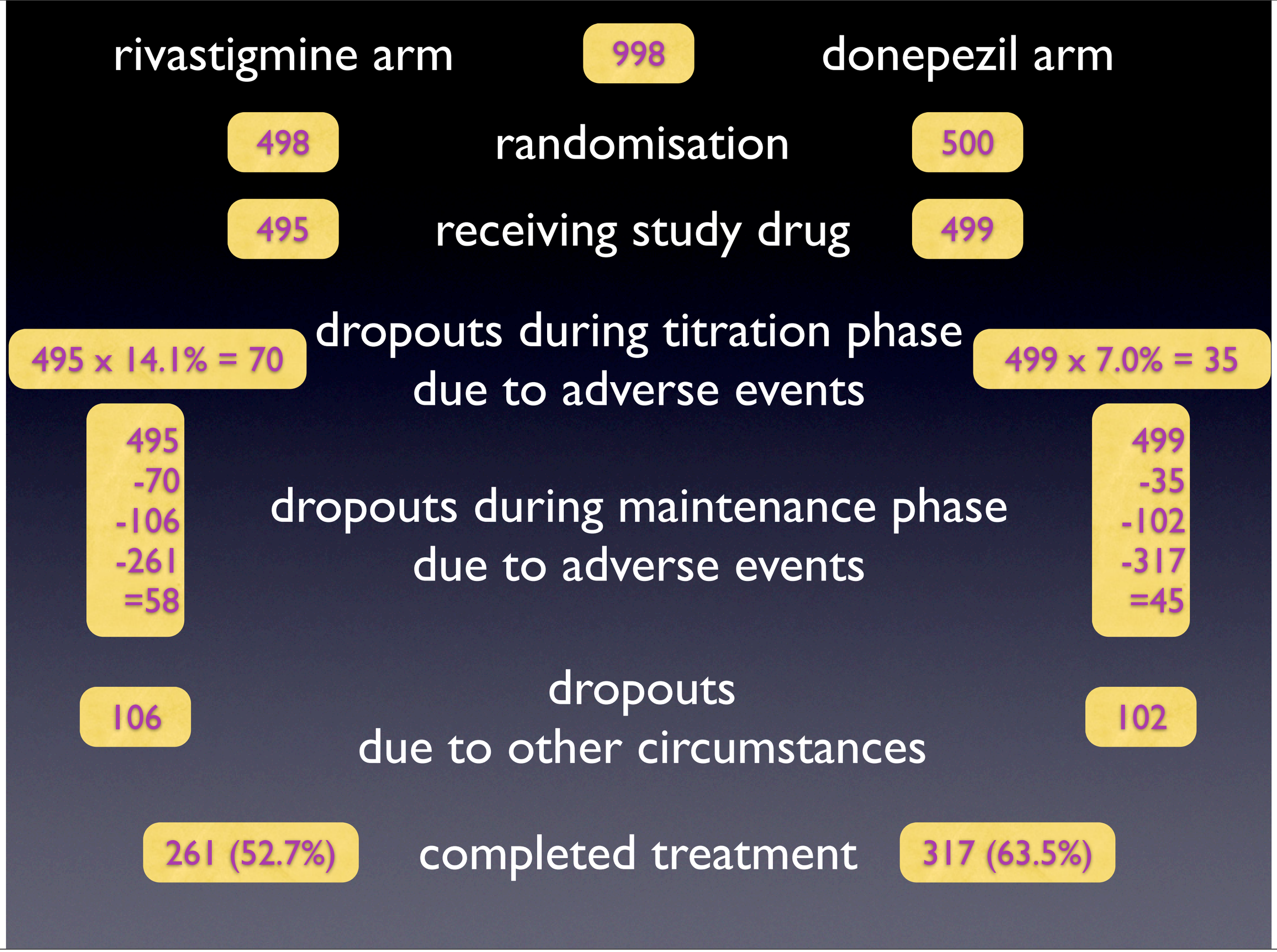
completed treatment

317 (63.5%)

in which only 35 patients dropped out.

This difference in dropout rates is important. It suggests that the rivastigmine is less well tolerated than the donepezil. Alternatively, the maximal doses for the two different drugs might reflect different amounts of active molecules. For example, what if the maximal dose for donepezil had been 20 mg per day, instead of 10? Would we then see equivalent rates of dropout due to adverse effects?

The actual number of dropouts during the maintenance phase can also be calculated (click on the 17.9% button)



So we get 58 dropouts due to adverse events in the maintenance phase for rivastigmine, and 45 for donepezil. These values are different but not as strikingly different as for the titration phase. This is likely because most of the people likely to experience intolerable adverse effects have already been weeded out during the titration phase.

This study used a technique called “Intention to treat” to deal with all these dropouts. In other words, anyone who started in the study and received the study drug is included in the final analysis, even if they dropped out during the study. Unfortunately, one may not have results of testing for the dropouts beyond the point at which they dropped out. Various approaches are used to fill in the missing test results. The approach used in this study is called “Last Observation Carried Forward” or LOCF for short. In this approach, one uses the last available test result for a dropout and carries forward that same value for the test result at the end of the study.

This type of analysis is extremely useful for many treatment trials in which patients tend to get better over time, even for placebo treatments. Typically, the last available test result for dropouts tends to reflect worse disease status than would a test result that was obtained for real at the end of the study. Thus, an ITT-LOCF approach tends to lowball the effect size of the treatment, compared to an approach in which only study completers are included in the analysis, especially if the treatment arm has more dropouts than the placebo arm.

However, this consideration may not apply when we are talking about a disease where the course tends to be downhill with or without the treatment. Now, using the last observation carried forward for dropouts means that an earlier test result, when the patient was less sick, would be used instead of a worse result. This wouldn’t matter much if the number of dropouts was the same for the two arms, but in the case of this study, where one arm had many more dropouts than the other, the result of using ITT-LOCF is that the arm with more dropouts looks like it is a more effective treatment!

Abbreviations

- SIB = Severe Impairment Battery
- s = received study drug (495 vs 499)
- c = completers (261 vs 317)
- t = adverse events dropouts during titration (70 vs 35)
- m = adverse events dropouts during maintenance (58 vs 45)
- z = other dropouts (106 vs 102)
- B = SIB mean score at baseline
- F = SIB mean score at week 104
- D = last SIB score for dropouts (LOCF)

For the study we are looking at, I have done a rudimentary analysis to see what the results might look like if, instead of an ITT-LOCF approach, a more traditional approach of only including patients who completed the study in the final result.

I will take you through this analysis step by step. As a start, here are the abbreviations I will be using.

F = B + delta

F = 87.77 - 9.30 = 78.47 (riva)

F = 87.80 - 9.91 = 77.89 (donep)

- SIB = Severe Impairment Battery
- s = received study drug (495 vs 499)
- c = completers (261 vs 317)
- t = adverse events dropouts during titration (70 vs 35)
- m = adverse events dropouts during maintenance (58 vs 45)
- z = other dropouts (106 vs 102)
- B = SIB mean score at baseline
- F = SIB mean score at week 104
- D = last SIB score for dropouts (LOCF)

First, let's calculate the actual final scores for the two arms, for the primary outcome measure, the Severe Impairment Battery. The final score F can be calculated as the initial score B, to which is added the amount of change, delta, over the 104 week study.

$$F_s.s = F_c.c + D_t.t + D_m.m + D_z.z$$

- SIB = Severe Impairment Battery
- s = received study drug (495 vs 499)
- c = completers (261 vs 317)
- t = adverse events dropouts during titration (70 vs 35)
- m = adverse events dropouts during maintenance (58 vs 45)
- z = other dropouts (106 vs 102)
- B = SIB mean score at baseline
- F = SIB mean score at week 104
- D = last SIB score for dropouts (LOCF)

This final score on the SIB, F, since it applies to all those who received the study drug, will have the subscript s. Similarly, the last SIB score for dropouts, D, which is the score carried forward, will have subscripts t, m, or z depending on whether it refers to adverse events dropouts during titration, adverse events dropouts during maintenance, or other dropouts, respectively.

Since all of these scores are average scores for the group, multiplying each average score by the number of patients in that group will give the total score for that group. And since the total final score for all the patients who received the study drug, $F_s.s$, must be the sum of the total scores for each of the groups of which the total group is composed, we arrive at the equation at the top of this slide.

$$(78.47)(495) = F_c.261 + D_t.70 + D_m.58 + D_z.106 \text{ (riva)}$$

- SIB = Severe Impairment Battery
- s = received study drug (495 vs 499)
- c = completers (261 vs 317)
- t = adverse events dropouts during titration (70 vs 35)
- m = adverse events dropouts during maintenance (58 vs 45)
- z = other dropouts (106 vs 102)
- B = SIB mean score at baseline
- F = SIB mean score at week 104
- D = last SIB score for dropouts (LOCF)

We have most of the numbers for this equation, for each of the study drug arms. Here, I've plugged in the numbers of patients in each group for rivastigmine,

$$(77.89)(499) = F_c.317 + D_t.35 + D_m.45 + D_z.102 \text{ (donep)}$$

- SIB = Severe Impairment Battery
- s = received study drug (495 vs 499)
- c = completers (261 vs 317)
- t = adverse events dropouts during titration (70 vs 35)
- m = adverse events dropouts during maintenance (58 vs 45)
- z = other dropouts (106 vs 102)
- B = SIB mean score at baseline
- F = SIB mean score at week 104
- D = last SIB score for dropouts (LOCF)

And here for the donepezil arm. Our goal is to calculate Fc, the SIB score for the group of patients who actually completed the study. This means we need to come up with values for Dt, Dm, and Dz, the mean scores for each of the dropout groups which could plausibly have been used in the study's calculations. I say, plausibly, because the paper does not actually give us these values.

Another complication is that the paper does not tell us how often the Severe Impairment Battery was applied to the patient group, other than saying "regularly". Does this mean weekly? Monthly? Every 6 months? The answer to this question may have a significant impact on the value for Fc that our equation will come up with.

simplifying assumptions

- actual numbers for results are lower than the numbers that received the study drug
  - rivastigmine: 471 instead of 495
  - donepezil: 483 instead of 499

- I will use the higher numbers (same as adding cases, using mean values for the rest of the group)

So we need to make some assumptions. The first arises because the table reporting the SIB results in the paper has n values which are less than the numbers that received the study drug. The paper does not explain how these numbers were arrived at. In any case, missing values can be dealt with by using the average value for the rest of the group, which is what in effect I will be doing by using the higher numbers of patients in my calculations.

simplifying assumptions

- linear decline in SIB scores
- constant rate of dropping out during each phase
- titration phase dropouts at week 8
- maintenance phase dropouts at week 16 + 88 / 2 = 60
- other dropouts at week 104 / 2 = 52

Another assumption has to do with the rate of decline in SIB scores. To simplify calculations, I have assumed that the rate of decline is a straight line from the starting point to the end point.

Without knowing when SIB assessments were done for each patient in the course of the study, it is difficult to even estimate what the last observation carried forward would be. I am making the assumption that SIB assessment results were continuously available, that the rate of dropping out remained constant for each of the dropout groups, and thus I can choose the midpoint of each phase as the point at which the average value for LOCF can be taken. These midpoints are week 8 for the 16-week titration phase, week 60 for the 88-week maintenance phase, and week 52 for the other dropouts. As the decline in SIB scores is assumed to be linear, the average SIB value at these midpoints is easy to calculate.

	riva	donep
B	87.77	87.80
delta	-9.30	-9.91
Dt = B + delta x 8 / 104	87.055	87.038
Dm = B + delta x 60 / 104	82.405	82.083
Dz = B + delta x 52 / 104	83.120	82.845

Using these midpoints, the last observation carried forward for each of the three groups works out to these values for the SIB.

$$(78.47)(495) = F_c.261 + D_t.70 + D_m.58 + D_z.106 \text{ (riva)}$$

$$(78.47)(495) = F_c.261 + (87.055)70 + (82.405)58 + (83.120)106$$

$$Fc = 73.405$$

Plugging these values into the equation I showed you earlier,

as I've done here,

gives the SIB score of 73.405 as the calculated final value for the group who completed the study in the rivastigmine arm.

$$(77.89)(499) = F_c.317 + D_t.35 + D_m.45 + D_z.102 \text{ (donep)}$$

$$(77.89)(499) = F_c.317 + (87.038)35 + (82.083)45 + (82.845)102$$

$$Fc = 74.690$$

Similarly, we arrive at a calculated final score of 74.690 for the donepezil patients who completed the study.

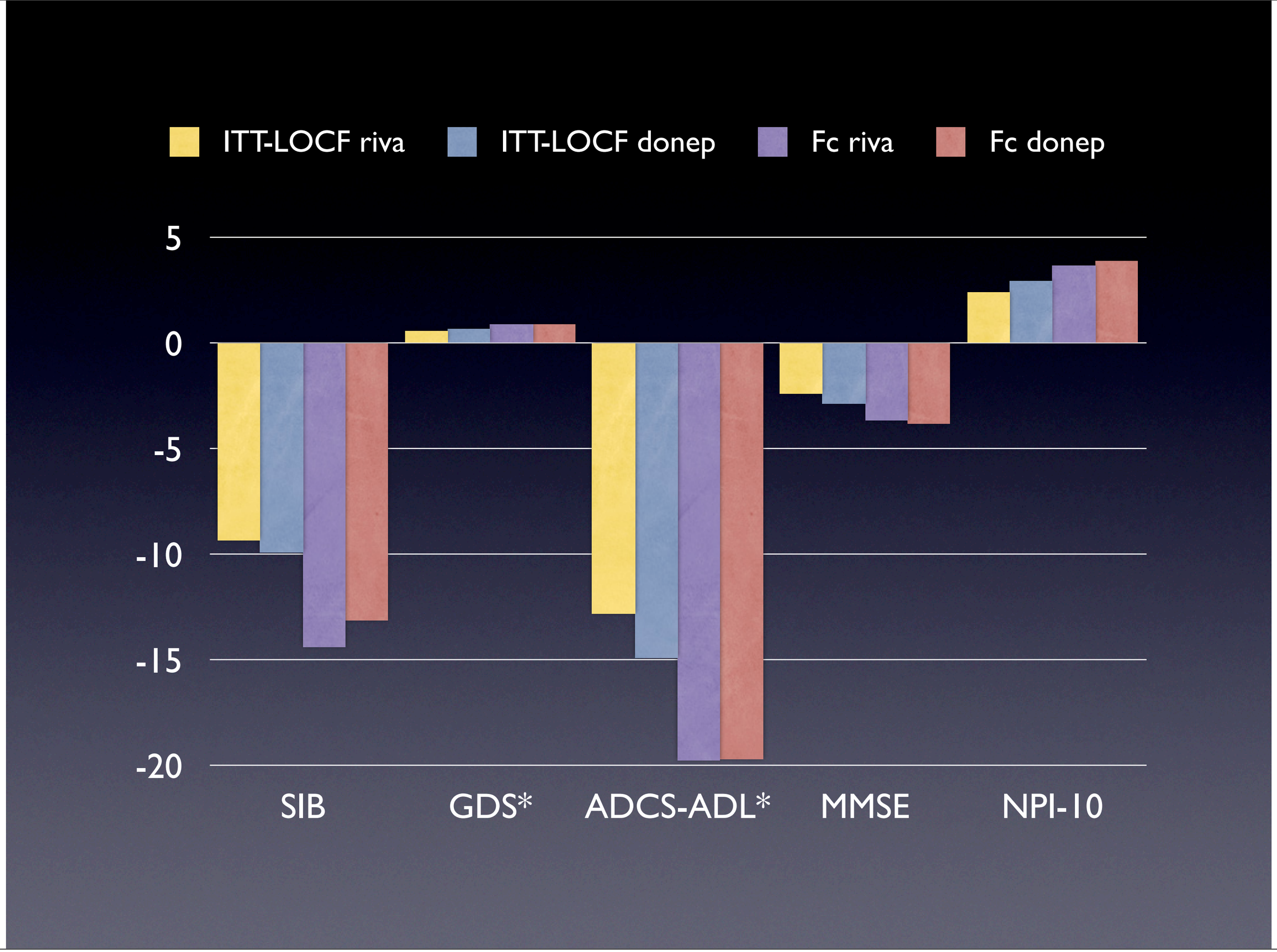
	ITT-LOCF			Fc		
	riva	donep		riva	donep	
SIB	-9.30	-9.91	0.61	-14.37	-13.11	-1.26
GDS	0.58	0.69	-0.11	0.90	0.91	-0.02
ADCS-ADL	-12.79	-14.87	2.08	-19.76	-19.67	-0.09
MMSE	-2.35	-2.85	0.50	-3.63	-3.77	0.14
NPI-10	2.40	2.94	-0.54	3.71	3.89	-0.18

This table shows the delta values for the various instruments used in the study, both the ITT-LOCF values from the paper, and the calculated values from my alternative analysis.

For the SIB, the ITT-LOCF analysis showed a decline for the rivastigmine group of 9.30, and 9.91 for the donepezil group. This difference in decline, favouring the rivastigmine group, was not statistically significant.

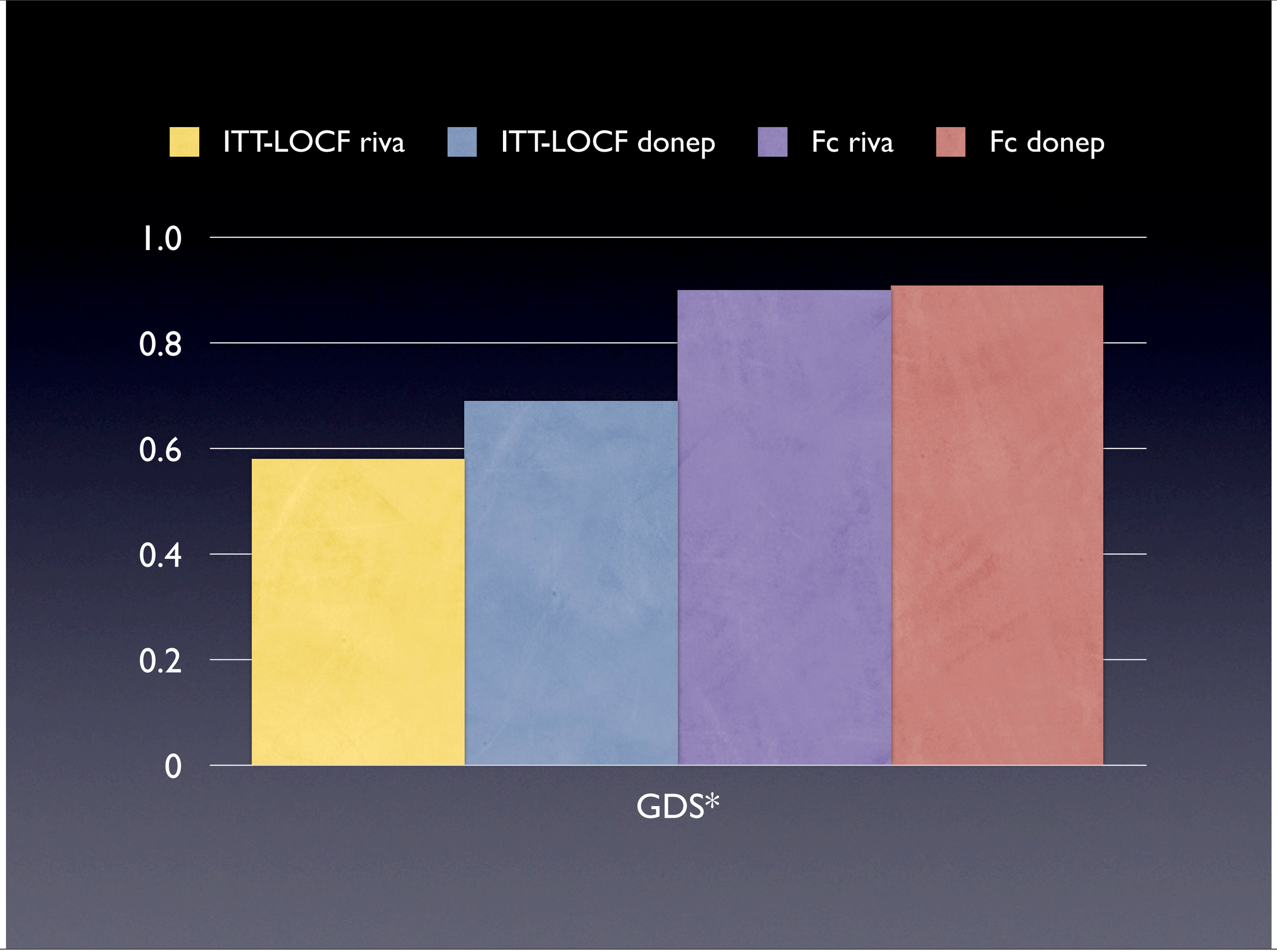
For my alternative analysis, it appears that the study completers in the rivastigmine arm would have had a decline of 14.37 on the SIB, whereas the donepezil completers would have declined by a little less, 13.11 over 2 years. This analysis then favours the donepezil group, although it is not possible to determine if the difference would be statistically significant.

The table also shows the values from the paper and my calculated values for the secondary outcome measures.

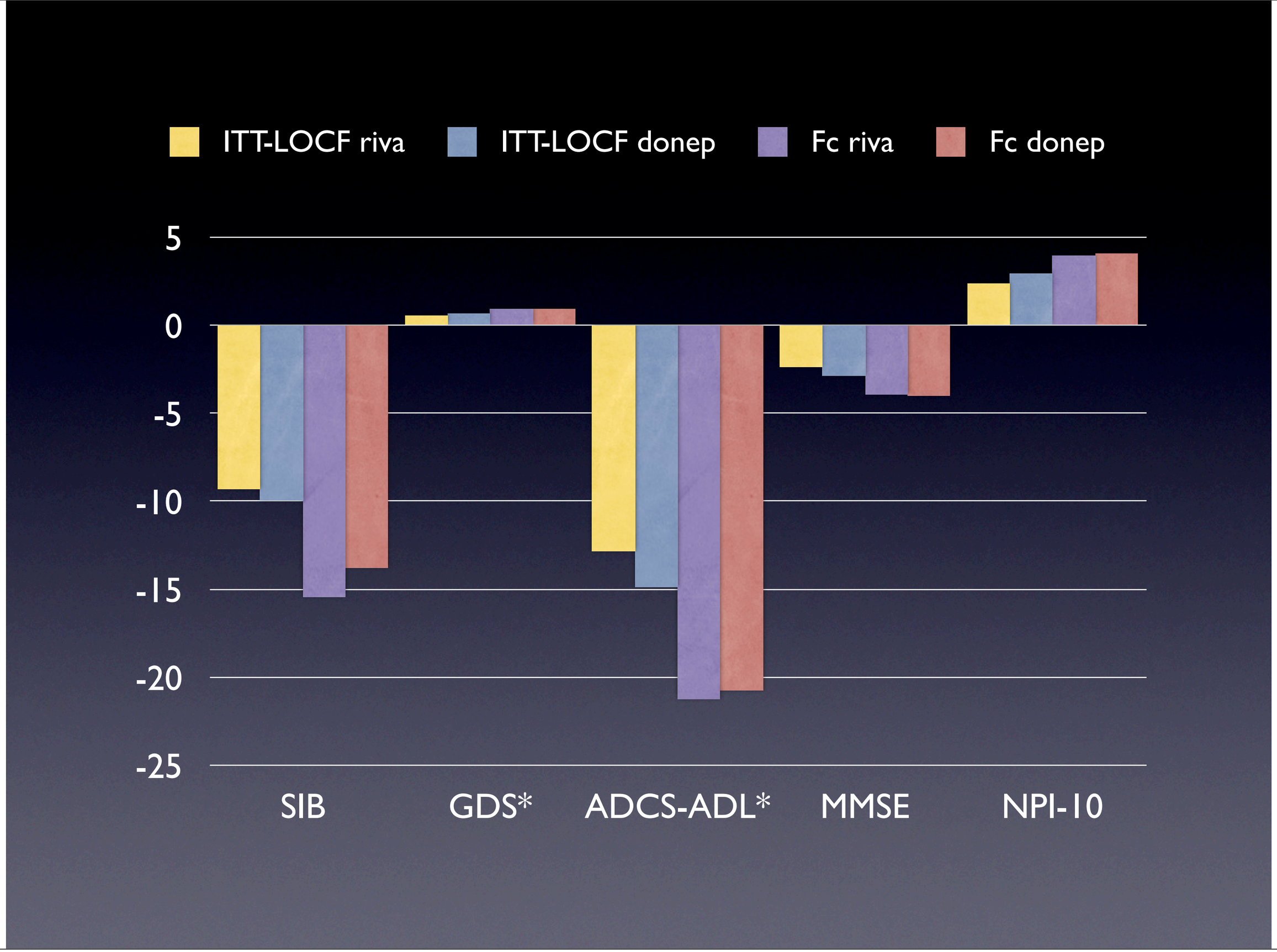


These are plotted here. Look at the middle set of bars, for the ADCS-ADL scores. In the paper, the authors reported that the decline in Activities of Daily Living scores was significantly less for rivastigmine than for donepezil. My alternative analysis shows that the declines would be identical for the two drugs.

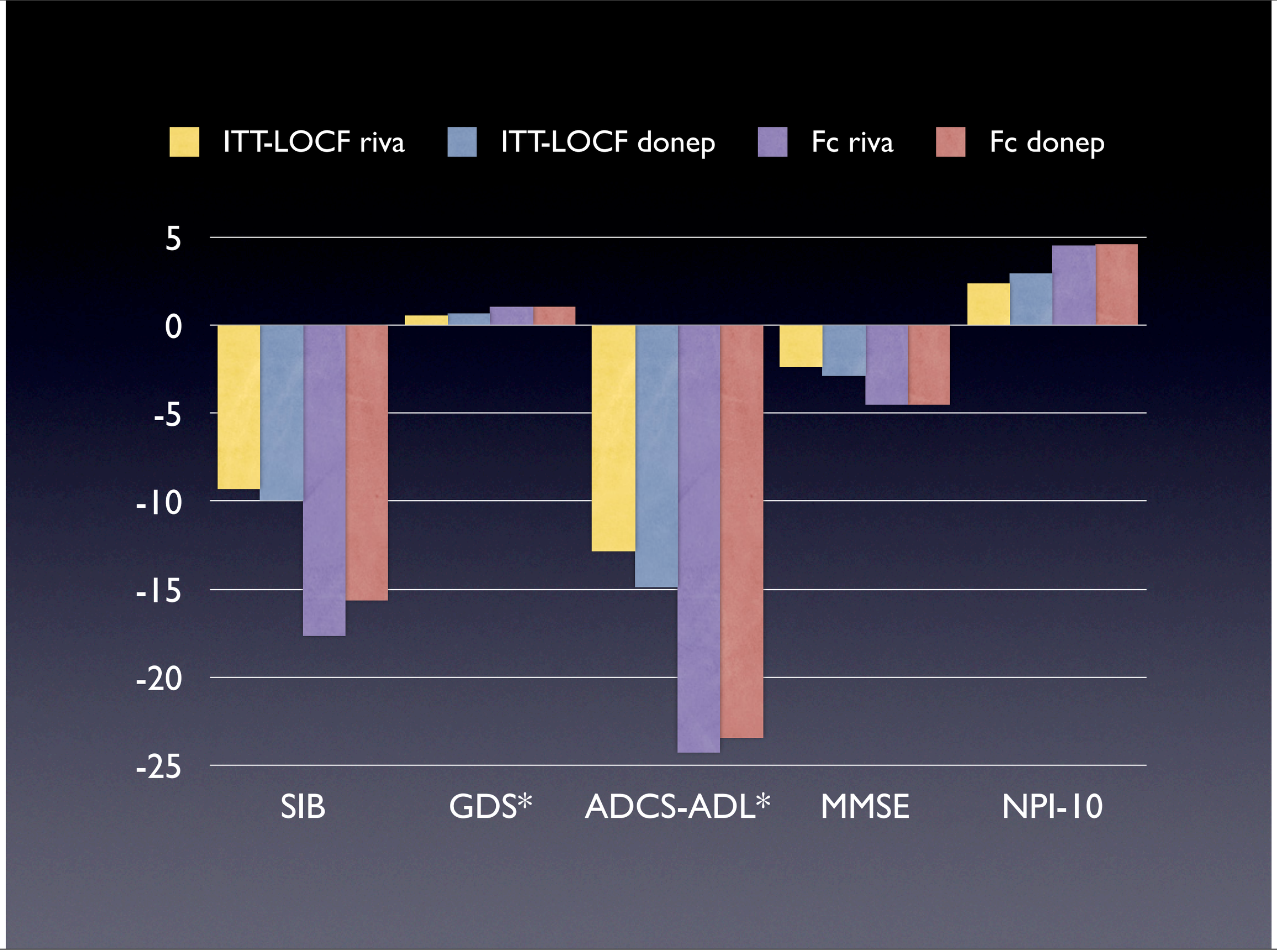
The set of bars for GDS, the Global Deterioration Score, are too small at this scale to read.



Here they are, blown up. Again, the paper reports significantly less deterioration for rivastigmine compared to donepezil, whereas my alternative analysis shows no difference in the amount of deterioration.



If we repeat the alternative analysis, but this time using the initial test results as the last observation carried forward for the adverse effects during titration dropouts, and calculated test results as of the beginning of the maintenance period for the adverse effects during maintenance dropouts, we begin to see a trend for the Activities of Daily Living scores in favour of donepezil.



Finally, if the various instruments were given only at the beginning and end of the study, then the LOCF values for all dropouts would default to their values at the beginning of the study. This is graphed here. One can see that the advantage in favour of donepezil is getting bigger, which demonstrates how important the LOCF strategy becomes when dropout rates are unequal.

Original conclusions

- Both drugs performed similarly on cognition and behavior
- rivastigmine may provide greater benefit in activities of daily living and global functioning

Revised conclusions

- There is a trend (possibly, significant difference) to less impairment in cognition with donepezil compared to rivastigmine over two years
- there was no difference between the two drugs on activities of daily living and global functioning

My point is that the conclusions one can draw from a study are in some cases quite dependent on the type of analysis that is used. In this study comparing two medications, paid for by the company that sells one of those medications, is it possible that the choice of analysis used in the published paper might have been influenced by the conclusions favouring the study sponsor's medication?

Because the two medications studied had quite different rates of side effects leading to important differences in dropout rates, and because the illness under study generally causes deterioration with or without treatment, an ITT-LOCF analysis is problematic. I would hope that the authors of the study will follow up with a publication of the actual results in their OC (observed case) population. As the authors themselves pointed out, carrying forward the last observation may lead to overestimations of drug effect, and there is a potential for bias related to the higher number of discontinuations in the rivastigmine population during the titration phase.