CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

761178Orig1s000

STATISTICAL REVIEW(S)

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Sciences
Office of Biostatistics

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/BLA #:	761178
Drug Name:	Aducanumab
Indication(s):	Alzheimer's
Applicant:	Biogen
Date(s):	July 7, 2020
Review Priority:	Priority
Biometrics Division:	I
Statistical Reviewer:	Tristan Massie, Ph.D.
Concurring Reviewers:	Kun Jin, Ph.D., Team Leader Sue-Jane Wang, Ph.D., Deputy Director James (Hsien-Ming) Hung, Ph.D., Division Director
Medical Division:	Division of Neurology I
Clinical Team:	Brian Trummer, M.D. Kevin Krudys, Ph.D. Ranjit Mani, M.D.
Project Manager:	Emilios (Andrew) Papanastasiou

Keywords: Futility Stopping, Substantial evidence

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 EXECUTIVE SUMMARY

Note that this statistical review focuses on the BLA dataset whereas the draft statistical review submitted as appendix 2 of the briefing package for the November 6, 2020 advisory committee meeting was based on the June 2019 dataset, which was the data evaluated by the collaborative Biogen/FDA workstream, since the June 2019 dataset was the first unblinded "final" dataset (there was only a difference in total CDRSB, primary endpoint, record counts of 4 out of 3716 total CDRSB records between the June 2019 and final BLA datasets [July 2020]). After the advisory committee the sponsor submitted comments on the draft statistical review at the time of the advisory committee (appendix 2 of the advisory committee briefing package). This statistical review contains our responses to the sponsor's comments in section 3.2.4 (page 86).

The two phase 3 studies (study 301, study 302) were stopped early for futility (March 21, 2019 press announcement) when both studies had reached 50% completion since it was estimated based on the interim study-pooled estimate of the treatment effects that both studies individually had <20% chance of success for either dose if completed. Following a futility press release announcement and collection of subsequent study closeout follow up data, the sponsor requested a meeting to discuss the two trials final data after discovering that despite the futility conclusion, the final analysis on face showed a statistically significant effect for the high dose in one of the two trials (p=0.01) but not the other (p=0.83).

Inconsistency on many levels summarizes the final clinical efficacy data from these trials. Because the two phase 3 studies were terminated for futility, the NDA package doesn't contain a single phase 3 study that was fully completed according to the plan. In fact, almost 50% are missing the Week 78 time point assessment of CDRSB which is the only timepoint that shows any significance and that is only significant in one of the two studies (the first study, study 301, high dose is numerically worse than placebo at Week 78 on the primary endpoint). A chance worse placebo response in study 302 than was observed in study 301 could explain the significance of study 302 (p=0.01).

This BLA submission does not have a situation such as just one study in existence and for which that study is strong. We have a second large adequate well controlled study that directly contradicts the first and is not even close to significance p=0.8252. If one has two studies and takes the best and pretends like it's the only study, one's estimate is most likely biased and misleading. In the opportunity to complete subset of 302 the high dose vs. placebo has a p-value of 0.0368 for the CDRSB at Week 78 and, even in the ITT population, there was no significance before Week 78. It is not justifiable to search for patients in 301 who are similar to 302 because that may have selection bias and presumes that 302 is right and 301 is wrong, for which there is no justification (without resorting to post-hoc analyses which are at best exploratory). Any selection of patients would need a proper placebo control. The overall 301 primary result is the

only valid well controlled, multiplicity adjusted, randomization validated analysis of 301 (and it had a substantial sample size).

The sponsor tries to discount study 301 due to post-hoc defined “rapid progressors”. Rapid progressors are likely part of the reality of Alzheimer’s and after the fact it is too late to address them in a completed large randomized study. Study 302 could just as well be the outlier relative to the true proportion of outliers in the natural progression. In fact, the range of CDRSB changes in Study 301 at 18 months appears consistent with the Alzheimer’s Disease Neuroimaging Study study (adnimerge_May15.2014 data). There are slightly more outliers in the high dose in 301, but that is worrisome in itself, since they are consistent with the ADNI data and so should again raise doubts about the representativeness of the 302 result. Furthermore, robust regression, techniques (M estimation , least trimmed squares, MM estimation, S estimation) designed to be resistant to and downweight outliers, applied to the 301 Week 78 data still suggest no effect of the high dose compared to placebo and that it was numerically worse than the low dose (vs. Placebo +0.0265 [S.E.=0.125], p=0.8315; vs. Low +0.0628 [S.E.=0.124], p=0.6153). Without the worst Week 78 CDRSB change of +13 in the high dose group the primary 301 high dose vs. placebo result is +0.0267 [S.E.=0.1495], p=0.8581 as compared to +0.0316 [S.E.=0.1499], p=0.8330 including it. This shows that Study 301 is a big study (the same size as study 302) and one outlier patient has limited influence. Totally excluding the patient instead of just the Week 78 observation the result is +0.0072 [S.E.=0.1487] still in the wrong direction for high dose vs placebo in study 301. Even excluding the 3 worst outliers for the high dose group, the high dose is still nowhere near significant in Study 301. More than one outlier in the high dose is more of a systemic problem and should be more worrisome and harder to discount. The sponsor also tries to use 301 to find a subgroup similar to Study 302, i.e., a subgroup showing efficacy in 301 but this relies on post-hoc non-randomized comparisons(Figure 14 on page 49). These analyses hide the fact that the post-hoc matched placebo progresses faster as the number of 10 mg/kg doses increases in these post-randomization event defined subgroups and such post-hoc matching can never equal a true randomization backed analysis. The only valid analysis of Study 301 is the prespecified randomization supported analysis of study 301 which failed for the high dose (p=0.83) and this study outcome should not be discounted without an extremely compelling reason (which there is not).

The sponsor argues, relying on non-randomized comparisons, that the high dose arm was challenged by intermediate dosing rather than full dosing in some patients. This can be countered by the fact that the low dose was numerically better than the high dose in Study 301, a comparison supported by randomization, and the low dose was also numerically better than high dose in study 302 in the subset after the mid-study protocol amendment increasing the maximum high dose for APOE carriers. Furthermore, the APOE non-carriers have less treatment effect on all four primary and key secondary efficacy endpoints despite having 10 mg/kg dosing from study start and less ARIA adverse events than APOE carriers, so fewer dose reductions due to

ARIA. In study 302 the estimated effect in APOE non-carriers on the primary endpoint is -.06 with 95%CI [-0.593, +0.471]. The high dose APOE non-carriers were also numerically worse than placebo in Study 302 on the first key secondary endpoint MMSE (treatment by APOE interaction term p=0.0096). In the APOE+ subgroup, which seems to drive study 302, the high dose was in the wrong direction overall in study 301 (APOE+: +0.07 [S.E.=0.18], p= 0.697). The study 302 success could be explained by a higher placebo progression after the implementation of protocol amendment 4 while the study was ongoing (Figure 6, page 38). This amendment increased the dose from 6 mg/kg to 10 mg/kg for APOE carriers, the stratum with more drug related ARIA adverse events with attendant individual patient dose titration modifications and unblinding (including some sponsor personnel) for the sake of dose managing (up to 35% of high dose patients had dose titration modifications). In 302 the occurrence of ARIA adverse events in the high dose was 1.4 times higher for APOE+ vs. APOE- prior to amendment 4 and 2.3 times higher post PV4. Limitation of dose titration in the high dose was 2.1 times higher for APOE+ prior to PV4 and 3.7 times higher post PV4. Thus, unblinding for dose managing may have been higher after PV4 (time to first ARIA in the high dose APOE+ subgroup also appears shorter after PV4). The APOE- stratum high dose had more 10 mg/kg doses as prescribed by the original protocol but was worse on average than APOE+ in 4 out of the 4 primary and key secondary endpoints (and the low dose shows the same pattern). This calls into question the sponsor's assertion about the importance of the actual number of 10 mg/kg doses received within the high dose group.

In the original "final" data presented to the Agency in June 2019, in Study 302 the MMSE had a p-value for the high dose of 0.0620 which would mean that no secondary endpoints in 302 would be significant following the prespecified hierarchy and multiplicity adjustment plan. In particular, the analysis plan suggests that the testing sequence was to compare all doses before moving to the next endpoint, which results in the same conclusion for secondary endpoints even after the final MMSE p-value decreased to 0.0493 for the high dose, because the low dose is not significant for any of the primary of key secondary endpoints (SAP excerpt: "for **each** of the secondary endpoints, a sequential (closed) testing procedure"). The sponsor has argued that the low dose effect is consistent across studies. However, the hierarchical multiplicity adjustment plan does not allow formally testing the low dose in study 301 since the high dose failed. Regardless, the low dose did not reach nominal significance in either study, so even if the effect was somewhat consistent it is not significantly different from placebo and has to be viewed in the context of the multiplicity of testing. These multiplicity considerations also highlight the issue that there were multiple final efficacy analyses and none of them at the sample size planned in the protocol, i.e., the sample size for Week 78 in the final analysis is not equal to the sample size of the futility analysis or the protocol planned maximum sample size. Thus, with an unplanned final sample size, the reported p-value is difficult to interpret in the usual frequentist sense that conceptualizes many repeated trials run according to the prespecified plan.

Given the large amount of missing data in the final ITT dataset (≥45% per group) and much lower rate missing in the Opportunity to Complete (OTC) dataset, some different demographics and disease characteristics in those without the opportunity to complete (due to futility stopping) that are related to outcome and time dependence of these not incorporated in the primary model (see discussion in section 3.2.1.4.2), the latter OTC dataset seems more relevant and reliable. The result for the Opportunity to Complete Dataset (total N=953) in Study 302 for the high dose difference from placebo on the primary endpoint, CDRSB at Week 78, was -0.36, 95% CI=[-0.70, -0.02], p=0.0368.

The primary objective of Study 103 was to evaluate safety and tolerability of multiple doses of Aducanumab in Alzheimer's patients. The study was exploratory and hypothesis generating for clinical efficacy. In particular, the statistical analysis plan designated the effect of aducanumab on clinical progression of AD as an exploratory objective after the primary objectives of safety and secondary objectives of cerebral amyloid plaque effects measured by 18F-AV-45 PET imaging, measurement of aducanumab in serum and evaluation of immunogenicity after multiple doses. The analysis plan further stated that "due to the exploratory nature of the study, there will be no multiple comparison adjustment". The sponsor's analysis of Study 103, 10 mg/kg arm vs. pooled placebo arms, is not supported by the randomization (3 of the placebo arms had no chance of receiving 10 mg/kg and one was entirely APOE carriers, while 10 mg/kg was not). Outside of rare diseases there is no justification for an analysis involving the pooling of staggered arms that is not supported by the study's overall randomization scheme. The comparisons that are supported by randomization (10 mg/kg [arm 4] vs. corresponding placebo [arm 5] p=0.12 and titration [arm 8] vs. corresponding placebo [arm 9] p=0.60 are not significant). A very small study without a proper randomization supported analysis should never have more weight than a much larger phase 3 randomized (parallel group) placebo controlled trial (e.g., Study 301).

In summary, the totality of the data does not seem to support the efficacy of the high dose. There is only one positive study at best and a second study which directly conflicts with the positive study. Both studies were not fully completed as they were terminated early for futility and had sporadic unblinding for dose management of ARIA cases which was much higher in the drug group(s). The Amyloid PET substudy data suggested a larger effect in APOE- (non-carriers) which is the opposite of what was observed for the overall clinical outcome data. Within the high dose group (or high and low combined) at the patient level there is no compelling correlation between the Week 78 change in the primary biomarker Aβ SUVR in the Composite region of interest with reference in the cerebellum and the Week 78 Change from baseline in CDRSB (see the biomarker section 3.2.1.4.2.2). For these reasons, substantial evidence has not been met in this application.

2 INTRODUCTION

2.1 Overview

The associated IND for the drug development was 106230. BIIB037 is a recombinant fully human antibody expressed in a CHO cell line, purified and formulated as a frozen liquid. BIIB037 is an IgG1 consisting of two heavy and two light chains connected by inter-chain disulfide bonds. BIIB037 has 1 carbohydrate moiety linked to Asn-304 in each heavy chain. The key studies intended to support efficacy are summarized in Table 1.

Table 1. Efficacy **Study Characteristics**

Study Name	Phase and Design	Treatment Period	# of Subjects, arms	Study Population
301	3 placebo controlled parallel study	78 Weeks	N=1647 total placebo/low/high APOE dependent high dose	MMSE 24-30 CDR global of 0.5 RBANS≤ 85
302	3 placebo controlled parallel study	78 Weeks	N=1638 total Placebo/low/high APOE dependent high dose	MMSE 24-30 CDR global of 0.5 RBANS≤ 85
103	1B Staggered Multiple Dose Design	54 Weeks	N=180 total 9 arms Arms 1-3: 1mg/kg: 3mg/kg: placebo Arms 4-5: 10 mg/kg /placebo 3:1 randomization Arms 6-7: 6 mg/kg:placebo 3:1 Arms 8-9: (Apoe+ only) titration to 10 mg/kg: placebo	Prodromal MMSE 24-30 Mild AD CDR global of 0.5 or 1.0 and MMSE 20-26

2.2 Data Sources

The primary efficacy ADAM and SDTM datasets for Study 302 were located in the following directory at the time of review.

\\cdsesub1\evsprod\bla761178\0003\m5\datasets\221ad302\analysis\adam\datasets\adqs.xpt

\\cdsesub1\evsprod\bla761178\0003\m5\datasets\221ad302\tabulations\sdtm\qs.xpt

The primary efficacy ADAM and SDTM data for Study 301 were located in the following directory at the time of review.
\\cdsesub1\evsprod\bla761178\0003\m5\datasets\221ad301\analysis\adam\datasets\adqs.xpt

\\cdsesub1\evsprod\bla761178\0003\m5\datasets\221ad301\tabulations\sdtm\qs.xpt

The primary efficacy data for Study 103 were located in the following directory at the time of review.

\\cdsesub1\evsprod\bla761178\0003\m5\datasets\221ad103\analysis\adam\datasets\adqs.xpt

\\cdsesub1\evsprod\bla761178\0003\m5\datasets\221ad103\tabulations\sdtm\qs.xpt

3 STATISTICAL EVALUATION

3.1 Data and Analysis Quality

It seems to be an uncommon situation for a sponsor to continue collecting and/or cleaning additional efficacy data a long time after most of data has been unblinded and analyzed (Study 302 BLA final data has 1581 subjects and 3716 post-baseline CDRSB records (877 have Week 78); June 2019 analysis data: has 1580 subjects and 3712 post-baseline CDRSB records [876 have Week 78 out of 1637 patients: 258 placebo and 276 high dose patients with the Opportunity

to Complete (OTC)]). [Study 301 BLA final data (July 2020) has 1602 patients and 3901 post-baseline CDRSB records; June 2019 analysis has 1602 patients and 3897 post-baseline CDRSB records]. In addition to the differences in record counts, some other CDRSB values were altered between the June 2019 and final BLA data submitted in July 2020. For example, for Week 78 records, 5 placebo, 4 low dose and 4 high dose CDRSB records were changed between the June 2019 dataset and the final BLA dataset (these corrections were implemented in the dataset after unblinding). The mean of these placebo Week 78 CDRSB records changed from 6.2 to 6.0, the mean for the low dose records changed from 1.83 to 0.63, and for the high dose from 0.50 to 1.25. Ten (10) placebo, 6 low dose, and 10 high dose patients changed from pre-PV4 amendment designation in the June 2019 data to post-PV4 in the final BLA data. With the addition of this data between June 2019 and the BLA submission the first key secondary, MMSE, went from non-significance to nominal significance in study 302 (p=0.0620 to 0.0493). The primary endpoint CDRSB in the final BLA data for the high dose minus placebo difference at Week 78 of -0.39, 95% CI=(-0.69,-0.09), p=0.0120 was more similar to the June 2019 data result (-0.40, p=0.0101), but the p-value for the low-placebo was slightly better in the final BLA data -0.26, (-0.57,+0.04), p=0.0901 as compared to June 2019: -0.25, p=0.1171.

Table 2 Differences between June 2019 dataset and final BLA dataset (July 2020)

		June 2019 dataset				BLA dataset			
Variable	Population	Placebo	Low	High	High-Placebo	Placebo	Low	High	High-Placebo
MMSE Week 78 LSMean	Overall	-3.26	-3.37	-2.72	.54 (.29) p=.0620	-3.26	-3.34	-2.69	.57 (.29) p=.0493
CDRSB Week 78 LSMean	Overall	1.74	1.49	1.34	-.40 p=.0101	1.74	1.48	1.35	-.39 p=.0120
	Pre-PV4	1.51	1.40	1.23	n/a	1.51	1.42	1.17	n/a
	Post-PV4	1.76	1.34	1.22	n/a	1.75	1.25	1.37	n/a
	APOE+ Pre-PV4	1.51	1.49	1.20	n/a	1.54	1.49	1.14	n/a
	APOE+ Post-PV4	2.23	1.30	1.28	n/a	2.13	1.22	1.42	n/a
Pre-PV4 subjects	Total P,L,&H=751					Total P,L,&H=722			
Post-PV4 subjects	Total P,L,&H=887					Total P,L,&H=916			
Post Baseline Starting of AD meds	10.8%					14.1%			

Notes:

- 1580 patients in June 2019 vs. 1581 in final BLA dataset (617-019 has a Week 26 and 50 record in final data only);

- 28 patients designated pre-pv4 in the June 2019 dataset changed to post pv4 designation based on a revised version 4 consent date in the final BLA dataset
- June:3712 CDRSB post-baseline visit records vs. BLA: 3716 post-baseline visit records.
- Revisions of June 2019 dataset records in the final BLA dataset: for Week 78, 5 placebo, 4 low dose and 4 high dose CDRSB records were revised

The remainder of this review focuses on the final BLA dataset except where noted.

3.2 Evaluation of Efficacy

3.2.1 Study 302

3.2.1.1 Study Design and Endpoints

Study 221AD302 [and similarly designed study 221AD301] was a multicenter, randomized, double-blind, placebo-controlled, parallel-group study in subjects with early Alzheimer's disease (AD), including mild cognitive impairment (MCI) due to AD and a subset of mild AD. Approximately 1605 subjects were to be enrolled across approximately 150 centers globally. The primary study objective is to evaluate the efficacy of monthly doses of aducanumab on the CDR-SB relative to placebo.

Subjects were to be randomized in a 1:1:1 ratio to 1 of the 3 treatment groups: aducanumab high dose, aducanumab low dose and placebo, with stratification based upon their apolipoprotein E4 (ApoE ε4) carrier status (carrier/non-carrier) and study site. During the placebo-controlled period, subjects were to receive infusions of aducanumab or placebo approximately every 4 weeks for approximately 18 months (a total of 20 doses). Dose levels were different in the same treatment group based upon subjects' ApoE ε4 carrier status, and specifically, ApoE ε4 carriers were to receive placebo, aducanumab 3 mg/kg, or aducanumab 10 mg/kg (note: this was 6 mg/kg before mid-study protocol amendment 4), whereas ApoE ε4 non-carriers were to receive placebo, aducanumab 6 mg/kg, or aducanumab 10 mg/kg.

Aducanumab was to be titrated for up to 6 doses prior to reaching the target. Note: As of Protocol Version 4 (mid-study implementation dated 24 March 2017), 10 mg/kg is the target dose for all ApoE ε4 carriers in the high-dose group. ApoE ε4 carriers who were randomized to the high dose group when the target dose was 6 mg/kg (under protocol versions prior to Version 4) must have received 2 or more doses at 6 mg/kg prior to being titrated up to 10 mg/kg. At the end

of the double-blind, placebo-controlled treatment period, subjects who met the extension entry criteria could enter a long-term safety and efficacy extension period, with all subjects receiving aducanumab approximately every 4 weeks (up to a total of 65 doses over 5 years).

Additionally, participants who developed ARIA (except those with asymptomatic, radiographically mild ARIA-H microhemorrhage) were to have a follow-up MRI performed every 4 weeks until the ARIA resolved (ARIA-E) or stabilized (ARIA-H), per the centrally read MRI. These participants were also to have MoCA assessments at these follow-up visits as well as biomarker, PK, and peripheral blood mononuclear cells samples collected at the first unscheduled visit following an episode of ARIA. Note: Participants with asymptomatic, mild ARIA-H microhemorrhages were exempt from these follow-up visits as mild ARIA-H microhemorrhage was observed at a similar incidence in aducanumab and placebo-treated participants in Study 221AD103.

The primary endpoint is the Change from baseline in CDRSB at Week 78.

Secondary endpoints have been rank prioritized, in the order shown below:

o Change from baseline in MMSE score at Week 78.

o Change from baseline in ADAS-Cog 13 score at Week 78.

o Change from baseline in ADCS-ADL-MCI score at Week 78.

3.2.1.2 Statistical Methodologies

Analysis Plan

Considerations for multiple comparison adjustments

A sequential (closed) testing procedure was to be used to control the overall Type I error rate due to multiple comparisons for the primary endpoint. The order of treatment comparisons was as follows: aducanumab high-dose versus placebo and aducanumab low-dose versus placebo. All comparisons after the initial comparison with $p > 0.05$ were not to be considered statistically significant.

Secondary endpoints have been rank prioritized, in the order shown

o Change from baseline in MMSE score at Week 78.

o Change from baseline in ADAS-Cog 13 score at Week 78.

o Change from baseline in ADCS-ADL-MCI score at Week 78.

Note that the key endpoints are also assessed at Week 26 and Week 50.

In order to control for a Type I error for the secondary endpoints, a sequential closed testing procedure was to be used and was to include both the order of the secondary endpoints and treatment comparisons. Specifically, for each of the secondary endpoints, a sequential (closed) testing procedure, as for the primary endpoint, was to be used to control the overall Type I error rate due to multiple treatment comparisons. If statistical significance is not achieved for 1 or 2 treatment comparisons, all endpoint(s) of a lower rank were not to be considered statistically significant for that 1 or 2 treatment comparisons, respectively.
*Reviewer's Comment: The closed testing for **each** of the secondary endpoints suggests that if the low dose is not significant the following tests and p-values for the high dose for lower endpoints in the hierarchy would not be allowable without inflating type I error. It seems that the plan was slightly ambiguous unless one interpreted it with the presumption that strong control of type I error over all key hypotheses is a requirement. The sponsor argued in their response to appendix 2 of the advisory committee briefing package that if the high dose was significant on the primary then it could be tested on the secondary regardless of the primary result for the low dose. However, the plan is not consistent with testing all key endpoints for the high dose before testing any key endpoints for the low dose. Type I error is not strongly controlled and could be as high as .0975 across all key hypotheses involving both doses and multiple endpoints if testing of the high dose could proceed regardless of the low dose result on the primary endpoint under the weak null, e.g., if the null hypothesis was false on the primary for one dose but true for the other dose and true for both doses on the secondaries then the chance of one or more type I errors in this scenario could be as high as .0975 if significance on the primary allowed further testing for the same dose regardless of the primary result for the other dose. For this reason strong control is needed.*

Primary analysis

The estimand of the primary analysis is the mean difference of the change from baseline CDR-SB scores at Week 78 between treatment groups in the ITT population [ICH E9 (R1) Addendum 2014, 2017]. All observed data was to be included in the primary analysis, including data collected after intercurrent events [ICH E9 (R1) Addendum 2017], i.e., treatment discontinuation or a change in concomitant use of AD symptomatic medication. The change from baseline CDR-SB scores was to be summarized by treatment group at each post-baseline visit. A mixed model repeated measures (MMRM) model was to be used as the primary analysis to analyze change from baseline CDR-SB using fixed effects of treatment group, time (categorical), treatment group-by-time interaction, baseline CDR-SB, baseline CDR-SB by time interaction, baseline MMSE, AD symptomatic medication use at baseline(yes/no), region, and laboratory ApoE ε4 status (carrier/non-carrier). An unstructured covariance matrix was to be used to model the within-patient variance-covariance errors. If the unstructured covariance structure matrix resulted

in a lack of convergence, the heterogeneous Toeplitz covariance structure followed by the heterogeneous first-order autoregressive covariance structure was to be used. The Kenward-Roger approximation was to be used to estimate the denominator degrees of freedom. In the primary analysis, missing data are assumed to be missing at random [Rubin 1976].

Sample Size Justification

A sample size of 450 subjects per treatment group (1350 in total) was planned to have approximately 90% power to detect a true mean difference of 0.5 in change from baseline CDR-SB at Week 78 between the 2 treatment groups. This power calculation was based on a 2-sided t-test assuming equal variance with a final significance level of 0.05, a standard deviation (SD) of 1.92 and a drop-out rate of 30%. The SD estimate of 1.92 for Week 78 reflected a 39% increase over the SD from the protocol-specified interim analysis of 1-year data. The assumed true mean difference of 0.5 between the 2 treatment groups represents an approximately 25% reduction in the placebo mean change from baseline at Week 78 if the placebo mean change is estimated to be 2. As defined in the protocol, the sample size for this study (and for the identically designed Study 221AD301) was reassessed in a blinded manner in November 2017 (approximately 3 months before enrollment completion and with about 10.6% of the data available on the primary endpoint from Studies 221AD301 and 221AD302 combined). At this timepoint, the SD of the primary endpoint was estimated based on the pooled blinded data from the two studies using a modified version of Gould-Shih simple-adjustment one sample variance (Zucker et al. 1999):

$$s_{adj}^2 = s_{os}^2 - \frac{2N}{9(N-1)}\delta^2,$$

where N denotes the number of subjects included in the analysis for blinded sample size reestimation (subjects with both baseline and Week 78 CDR-SB available at the time of sample size re-estimation), δ is the assumed true treatment effect (same treatment effect assumed for both the high dose group and low dose group in this analysis), and S_{OS}^2 is the unadjusted one sample variance of the primary endpoint estimate from the pooled blinded data. As a result of this analysis, the sample size was increased from 1350 to 1605 (450 to 535 per treatment group) to assure adequate power for detecting a mean treatment effect of 0.5.

Interim Analysis

An interim analysis was planned to occur after approximately 50% of the subjects had the opportunity to complete the Week 78 visit for both 221AD301 and 221AD302. To maintain the integrity of the study in the event of the interim analysis, an independent group external to

Biogen, that was not to be involved in the conduct of the study after unblinding, was to perform the interim analysis. The IDMC was to review the unblinded results of the interim analysis provided by the independent group and was to make a recommendation to Biogen based on pre-specified criteria.

An interim analysis for futility of the primary endpoint was to be performed to allow early termination of the studies if it was evident that the efficacy of aducanumab was unlikely to be achieved. The futility criteria were to be based on conditional power, which is the chance that the primary efficacy endpoint analysis will be statistically significant in favor of aducanumab at the planned final analysis, given the data at the interim analysis. The conditional power is calculated assuming that the future unobserved effect is equal to the maximum likelihood estimate of what is observed in the interim data:

$$CP\big(Z(1) \geq Z_\alpha \big| Z(t)\big) = 1 - \phi\left(\frac{Z_\alpha \sqrt{n_2} - Z(t)\sqrt{n_1}}{\sqrt{n_2 - n_1}} - \frac{Z(t)\sqrt{n_2 - n_1}}{\sqrt{n_1}}\right)$$

where t is the fraction of information and $Z(t)$ is the observed Z-statistic at the interim analysis, $Z(1)$ is the Z-statistic and α is the type I error at the final analysis, n_1 and n_2 are the numbers of subjects at the interim and at the final analysis, respectively.

The futility decision was to primarily be based on the conditional power for the primary efficacy endpoint. The study was not to be considered as futile unless both studies 221AD301 and 221AD302 had conditional power for the primary efficacy endpoint less than 20% in both the high-dose and low-dose treatment groups. Given the insufficient knowledge of aducanumab's potential effects on various functional/cognition endpoints or in certain subgroups at the planning time, other data in addition to the pre-specified futility criteria was to be considered as well, and the IDMC may have recommended the studies to be continued as planned based on the weight of the evidence.

The Statistical analysis plan stated that an interim analysis for superiority may be performed, to allow the possibility to demonstrate the treatment effect early. If an interim analysis for superiority was performed, the O'Brien-Fleming stopping boundary was to be used. If an interim analysis for superiority was not performed, then no alpha adjustment would be used for the final analysis after all subjects have had the chance to complete the Week 78 visit. The SAP provided no other details on this interim efficacy analysis.

Reviewer's Comment: The Statistical analysis plan and the Unblinding plan do not definitively state whether the interim (futility) analysis was to include data from ongoing subjects who had not had the opportunity to complete Week 78.

Responder analysis

To further assess whether subjects on aducanumab progress differently from those on placebo, responder analysis was to be conducted. The responders were to be determined by a threshold of the primary endpoint, i.e., subjects whose change from baseline CDR-SB at Week 78 is smaller than or equal to the threshold were to be classified as responders and otherwise were to be classified as non-responders. All subjects with missing data at Week 78 were to be classified as non-responders.

The responder analysis was to be conducted for two threshold values: 0.5 or 1.5, i.e., subjects whose change from baseline CDR-SB at Week 78 ≤ 0.5 or ≤ 1.5. The number of responders and the response rate were to be summarized by treatment group. The dichotomized response, responder vs. non-responder, were to be modeled using a logistic regression with the following covariates: treatment group, baseline CDR-SB, baseline MMSE, AD symptomatic medication use at baseline (yes/no), region, and laboratory ApoE ε4 status (carrier/non-carrier). In addition to the two selected threshold values, the continuous responder curve that displays the percentage of responders under a wide range of threshold values was to be presented by treatment group.

Amyloid PET Analysis

Amyloid PET substudy

Every subject enrolled into the study must have a positive amyloid PET scan by visual read either at screening or obtained within 12 months of screening. Subjects enrolled into the amyloid PET substudy will have the quantitative standard uptake value ratio (SUVR) scores at screening and at each planned post-baseline visit. The amyloid PET substudy was to include a subset of approximately 400 subjects in countries other than Japan where PET scans were to be performed using ^{18}F-florbetapir ligand, and a small subset of subjects in Japan where either ^{18}F-florbetapir ligand or ^{18}F-flutemetamol ligand could be used. In the placebo-controlled period, amyloid PET assessments were scheduled at screening, Week 26, and Week 78.

Amyloid PET SUVR regions-of-interest and reference regions

Amyloid PET standardized uptake value ratio (SUVR) is a quantitative measure of cerebral amyloid plaque burden. The SUVR was to be calculated for the following target brain regions of-interest (ROIs): composite ROI, frontal cortex, parietal cortex, lateral temporal cortex, sensorimotor cortex, anterior cingulate cortex, posterior cingulate cortex, medial temporal cortex, occipital cortex, striatum, and statistical ROI normalized to reference region activity. Additionally, SUVR ROIs including pons and deep subcortical white matter which are believed to be least affected by amyloid pathology were also to be evaluated. The composite ROI was to be comprised of major cortical regions part of the frontal, parietal, lateral temporal, sensorimotor, anterior, posterior cingulate and occipital cortices to serve as a summary measure of global cerebral amyloid burden. The statistical ROI is a region of interest consisting of the posterior cingulate cortex, precuneus and medial frontal cortex that has been demonstrated to yield optimal group separation between subjects with low and high amyloid burden across different reference regions. A negative change from baseline in composite ROI SUVR indicates

a reduction in amyloid burden and a negative treatment difference (aducanumab minus placebo) favors aducanumab. The composite ROI was to serve as the ROI of primary focus. The following reference regions were to be employed: cerebellum, cerebellum cropped, cerebellar white matter, cerebellar grey matter, deep subcortical white matter, pons, cerebellum + pons, cerebellar white matter + pons, deep subcortical white matter + cerebellum, deep subcortical white matter + pons and deep subcortical white matter + cerebellum + pons. Cerebellum was to serve as the reference region of primary focus. The composite ROI SUVR using cerebellum as the reference region was to be used as the primary endpoint for amyloid PET analysis.

Amyloid PET analysis population

There are two amyloid PET analysis populations: 18F-florbetapir amyloid PET analysis population and 18F-flutemetamol amyloid PET analysis population.

By Visit summary and MMRM model

The baseline and change from baseline amyloid PET SUVR values were to be summarized by treatment groups (placebo, low dose and high dose) and by visit for each of the target ROIs using cerebellum as the reference region for each of the amyloid PET analysis populations. In addition, the baseline and change from baseline amyloid composite ROI values were to be summarized by treatment groups by visit for each of the reference regions for each of the amyloid PET analysis populations.

For the ^{18}F-florbetapir amyloid PET analysis population, an MMRM model was to be used to analyze change from baseline SUVR for each target ROI with cerebellum as the reference region. Fixed effects of the model were to include treatment groups (placebo, low dose and high dose), visit (Week 26 and Week 78), treatment group-by-visit interaction, baseline SUVR (continuous), baseline SUVR by visit interaction, baseline MMSE (continuous), laboratory ApoE ε4 status (carrier and non-carrier), and baseline age (continuous). Visit and treatment group were to be treated as categorical variables in the model along with their interactions. An unstructured covariance matrix was to be used to model the within-patient variance-covariance errors. If the unstructured covariance structure matrix resulted in a lack of convergence in any of the parameters, the heterogeneous Toeplitz covariance structure followed by the heterogeneous first-order autoregressive covariance structure was to be used for all the parameters. The Kenward-Roger approximation was to be used to estimate the denominator degrees of freedom. Adjusted means for each treatment group, pairwise adjusted differences with placebo, 95% confidence intervals for the differences and associated p-values were to be presented at week 26 and week 78. The same MMRM model was also to be used to analyze the change from baseline SUVR for the composite ROI with each of the reference regions. No multiple comparison adjustment was to be used for amyloid PET analysis.

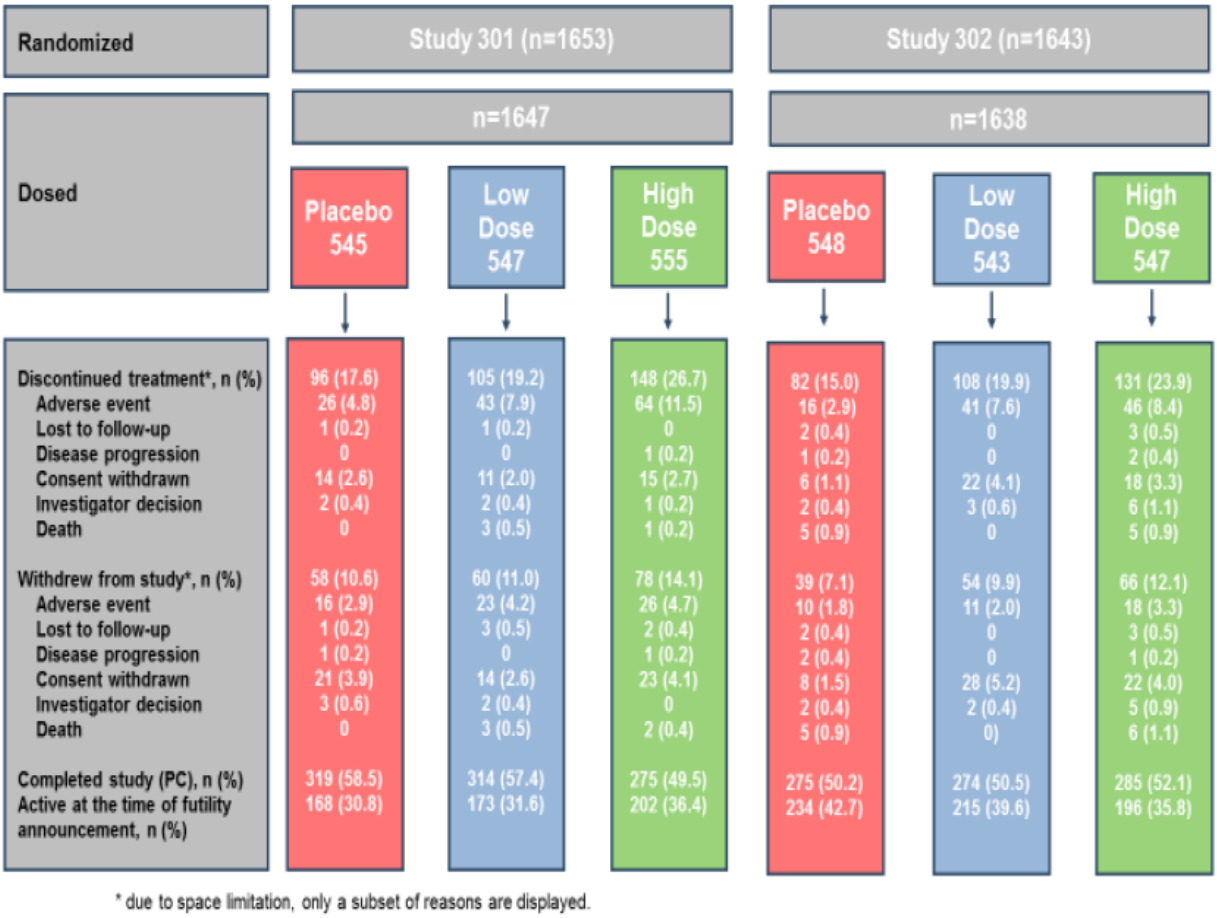
3.2.1.3 Patient Disposition, Demographic and Baseline Characteristics

Subject Accountability

Subject flow through the study is shown in Table 3. Patient Disposition. In both studies, incidence of discontinuation due to AEs was highest in the aducanumab high dose group.

Reviewer's Comment: The High dose had a larger proportion of patients discontinuing treatment, mostly due to AEs. This might affect the efficacy assessment because the missing data may be missing not at random, i.e., worse than completers.

Table 3. Patient Disposition in the Phase 3 Trials



Data source: t-acct-sub-pc-new-301/Output 5, t-acct-sub-pc-new-302/Output 6

Note: This table was copied from page 18 of the sponsor's 6/14/19 briefing package

Baseline Disease Characteristics

Subject demographics (Table 4) and baseline disease characteristics (Table 5) were balanced across groups in both phase 3 Studies.

Table 4. Baseline Demographics for Studies 301 and 302

	Study 301			Study 302		
Dosed	**Placebo (n=545)**	**Low Dose (n=547)**	**High Dose (n=555)**	**Placebo (n=548)**	**Low Dose (n=543)**	**High Dose (n=547)**
Age in years, mean ± SD	69.8±7.72	70.4±6.96	70.0±7.65	70.8±7.40	70.6±7.45	70.6±7.47
Gender, Female n (%)	287 (52.7)	284 (51.9)	292 (52.6)	290 (52.9)	269 (49.5)	284 (51.9)
Race						
Asian n (%)	55 (10.1)	55 (10.1)	65 (11.7)	47 (8.6)	38 (7.0)	41 (7.5)
White n (%)	413 (75.8)	412 (75.3)	413 (74.4)	415 (75.7)	418 (77.0)	405 (74.0)
Education years, mean ± SD	14.7±3.66	14.6±3.77	14.6±3.72	14.5±3.82	14.5±3.63	14.6±3.74
AD medications used, n (%)	293 (53.8)	307 (56.1)	307 (55.3)	279 (50.9)	277 (51.0)	277 (50.6)
ApoE ε4, n (%)						
Carriers	376 (69.0)	391 (71.5)	378 (68.1)	367 (67.0)	362 (66.7)	365 (66.7)
Homozygote E4	104 (19.1)	101 (18.5)	104 (18.7)	92 (16.8)	97 (17.9)	77 (14.1)
Heterozygote E4	272 (49.9)	290 (53.0)	274 (49.4)	275 (50.2)	265 (48.8)	288 (52.7)
Non-carriers	167 (30.6)	156 (28.5)	176 (31.7)	178 (32.5)	178 (32.8)	181 (33.1)
Clinical stage, n (%)						
MCI due to AD	443 (81.3)	440 (80.4)	442 (79.6)	446 (81.4)	452 (83.2)	438 (80.1)
Mild AD	102 (18.7)	107 (19.6)	113 (20.4)	102(18.6)	91 (16.8)	109 (19.9)
PET SUVR, mean composite ± SD (n) – **PET substudy only**	1.38±0.198 (203)	1.39±0.186 (198)	1.41±0.177 (181)	1.37±0.175 (157)	1.39±0.181 (157)	1.38±0.183 (171)

Note: Table Copied from page 19 of 6/14/19 briefing package

Table 5. Baseline Disease Characteristics for Studies 301 and 302

Dosed	Study 301			Study 302		
	Placebo (n=545)	Low Dose (n=547)	High Dose (n=555)	Placebo (n=548)	Low Dose (n=543)	High Dose (n=547)
RBANS delayed memory score, mean ± SD	60.0±13.65	59.5±14.16	60.6±14.09	60.5±14.23	60.0±14.02	60.7±14.15
MMSE, mean ± SD	26.4±1.73	26.4±1.78	26.4±1.77	26.4±1.78	26.3±1.72	26.3±1.68
CDR Global Score, n (%) 0.5 1	 544 (99.8) 1 (0.2)	 546 (99.8) 1 (0.2)	 554 (99.8) 0	 544 (99.3) 3 (0.5)	 543 (100) 0	 546 (99.8) 1 (0.2)
CDR-SB, mean ± SD	2.40±1.012	2.43±1.014	2.40±1.009	2.47±0.999	2.46±1.011	2.51±1.053
CDR cognitive subscore, mean ± SD	1.73±0.623	1.75±0.615	1.72±0.612	1.75±0.644	1.76±0.643	1.78±0.650
CDR functional subscore, mean ± SD	0.68±0.574	0.68±0.558	0.68±0.585	0.72±0.554	0.71±0.558	0.73±0.577
ADAS-Cog 13, mean ± SD	22.5±6.56	22.5±6.30	22.4±6.54	21.9±6.73	22.5±6.76	22.2±7.08
ADCS-ADL-MCI score, mean ± SD	43.0±5.55	42.9±5.73	42.9±5.70	42.6±5.73	42.8±5.48	42.5±5.82

Data source: t-bl-char-pc-new-301/Output 7, t-bl-char-pc-new-302/Output 8

Note: Table Copied from page 20 of 6/14/19 briefing package

3.2.1.4 Results and Conclusions

3.2.1.4.1 Sponsor's Results

Table 6 shows the sponsor's results for both Study 301, Study 302 and a pooled study analysis by the sponsor presented to the Division at a Type C meeting in 2019. The sponsor's final analysis of the first key secondary endpoint, MMSE, had a p-value of 0.0493 (down from .0620 as presented to the Agency in June 2019) after the addition of a few more records and a small proportion of revisions that were collected after unblinding (Table 7).

Table 6. Phase 3 Primary and Key Secondary Results (June 2019 results)

	Study 301			Study 302			Studies 301+302		
		Diff vs PBO [a] (%) p-value			Diff vs PBO [a] (%) p-value			Diff vs PBO [a] (%) p-value	
	PBO decline (n=545)	Low dose (N=547)	High dose (N=555)	PBO decline (n=548)	Low dose (N=543)	High dose (N=547)	PBO decline (n=1093)	Low dose (N=1090)	High dose (N=1102)
CDR-SB	1.55	-0.18 (-12%) 0.2362	0.03 (2%) 0.8252	1.74	-0.25 (-14%) 0.1171	-0.40 (-23%) 0.0101	1.64	-0.21 (-13%) 0.0513	-0.18 (-11%) 0.0974
MMSE	-3.5	0.2 (-6%) 0.4875	-0.1 (3%) 0.7961	-3.3	-0.1 (3%) 0.6900	0.5 (-15%) 0.0620	-3.4	0 (0%) 0.8346	0.2 (-6%) 0.2621
ADAS-Cog13	5.171	-0.590 (-11%) 0.2475	-0.605 (-12%) 0.2446	5.171	-0.747 (-14%) 0.1672	-1.395 (-27%) 0.0098	5.171	-0.657 (-13%) 0.0764	-0.989 (-19%) 0.0083
ADCS-ADL-MCI	-3.8	0.7 (-18%) 0.1345	0.7 (-18%) 0.1520	-4.3	0.7 (-16%) 0.1556	1.7 (-40%) 0.0009	-4.0	0.7 (-18%) 0.0347	1.2 (-30%) 0.0005

[a]: difference vs placebo at week 78. Negative percentage means less progression in the treated arm.

N: numbers of all randomized and dosed subjects that were included in the ITT analysis.

Note: This table was copied from page 23 of the 6/14/19 briefing package

Table 7 Sponsor's Final BLA data phase 3 Results for Primary and Key Secondary endpoints

301

		Diff vs PBO [a] (%) p-value	
	PBO decline (N=545)	Low dose (N=547)	High dose (N=555)
CDR-SB	1.56	-0.18 (-12%) 0.2250	0.03 (2%) 0.8330
MMSE	-3.5	0.2 -6% 0.4795	-0.1 3% 0.8106
ADAS-Cog13	5.140	-0.583 -11% 0.2536	-0.588 -11% 0.2578
ADCS-ADL-MCI	-3.8	0.7 -18% 0.1225	0.7 -18% 0.1506

[a]: difference vs placebo at Week 78. Negative percentage means less progression in the treated arm.

N: numbers of all randomized and dosed participants that were included in the ITT analysis.

Note: The ITT analysis was conducted on the ITT population excluding data collected after 20 March 2019.

Note: A mixed model repeated measures (MMRM) was used as the primary analysis to analyze change from baseline CDR-SB using fixed effects of treatment group, time (categorical), treatment group-by-time interaction, baseline CDR-SB, baseline CDR-SB by time interaction, baseline MMSE, Alzheimer's disease symptomatic medication use at baseline, region, and laboratory ApoE e4 status. The change from baseline in MMSE, ADAS-Cog13 and ADCS-ADL-MCI scores were also analyzed using MMRM.

Data source: 2.7.3 Summary of Clinical Efficacy, Table 8 and Table 9

302

		Diff vs PBO [a] (%) p-value	
	PBO decline (N=548)	Low dose (N=543)	High dose (N=547)
CDR-SB	1.74	-0.26 -15% 0.0901	-0.39 -22% 0.0120
MMSE	-3.3	-0.1 3% 0.7578	0.6 -18% 0.0493
ADAS-Cog13	5.162	-0.701 -14% 0.1962	-1.400 -27% 0.0097
ADCS-ADL-MCI	-4.3	0.7 -16% 0.1515	1.7 -40% 0.0006

[a]: difference vs placebo at Week 78. Negative percentage means less progression in the treated arm.

N: numbers of all randomized and dosed participants that were included in the ITT analysis.

Note: The ITT analysis was conducted on the ITT population excluding data collected after 20 March 2019.

Note: A mixed model repeated measures (MMRM) was used as the primary analysis to analyze change from baseline in the CDR-SB using fixed effects of treatment group, time (categorical), treatment group-by-time interaction, baseline CDR-SB, baseline CDR-SB by time interaction, baseline MMSE, Alzheimer's disease symptomatic medication use at baseline, region, and laboratory ApoE e4 status. The change from baseline in MMSE, ADAS-Cog13 and ADCS-ADL-MCI scores were also analyzed using MMRM.

Data source: 2.7.3 Summary of Clinical Efficacy Table 2 and Table 3

Note: Sponsor's final BLA dataset results copied from pages 43 and 48 of sponsor's clinical overview.

Reviewer's Comment: Week 78 High vs Placebo CDR-SB 95% Confidence Intervals: 302: -0.390 [S.E.=0.155] (-0.694, -0.086); 301: +0.032 [S.E.=0.150] (-0.262, 0.326). The Placebo LSMean at Week 78 was 1.56 and 1.74 in 301 and 302, respectively, and Aducanumab High LSMean was 1.59 and 1.35.

A large proportion of the ITT population (~45%) did not have the opportunity to complete Week 78 due to the futility stopping of the trials. The result for the Opportunity to Complete Dataset (N=953) in Study 302 for the high dose on the primary endpoint, CDRSB at Week 78, was -0.36, 95% CI = [-0.70, 0-.02], p=0.0368.

Reviewer's Comment: The sponsor's highlighting of percent reduction from placebo masks the placebo effect which is highly variable and doesn't represent the analysis scale or acknowledge the standard error of the percent reduction which is needed for proper context. The p-values reflect the primary analysis model which evaluated the simple difference μ_d - $\mu_{p,}$ not the percent difference. Percent reduction should really be estimated using a different model, e.g., log(CDRSB), and p-values and standard errors would be different, thus the percent reduction should be interpreted with caution.

3.2.1.4.2 Reviewer's Results

3.2.1.4.2.1 Primary and Sensitivity Analyses

Note that the sponsor published the placebo controlled period results of study 103 in September 2016 and presented the 48 month analysis of study 103 results at an Alzheimer's meeting in October 2018. The date of first treatment in study 302 was 15 September 2015.

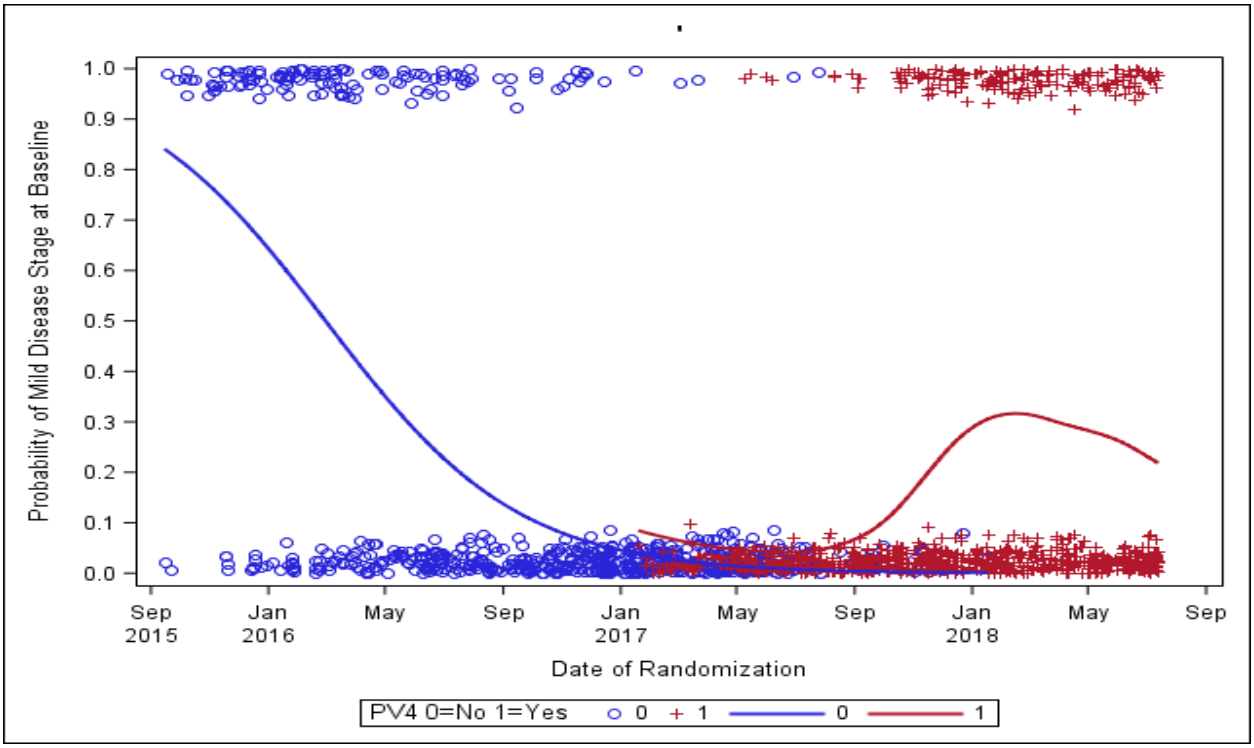
There is a lot of missing data in study 302 (and 301) at Week 78 (>40%) caused by early stopping due to futility. Week 78 is the only Visit with apparent efficacy for CDRSB in study 302. Unblinding due to ARIA and high dose titration limitations vary significantly by APOE+ vs. APOE- (as do, to some extent, CDRSB changes) and the maximum dose for the high dose APOE+ changed after protocol amendment 4. Therefore, Missing at Random may not be a reasonable assumption for missing data without stratifying by APOE.

Those missing Week 78 due to late enrollment and early stopping have some differences in baseline characteristics compared to those who had the opportunity to complete: Baseline use of symptomatic medications overall was increased by 6.7%, Region 1 decreased from 46 to 29% (region 2 increased from 50 to 57% and region 3 increased from 3 to 13%), the percentage Mild diagnosis increased from 15 to 25%, baseline CDRSB for the high dose is about 0.20 points higher and 0.05 lower for placebo, and the oldest age group proportion increased by 5% after Sep 21, 2017 as compared to those randomized before the same date (i.e., with the opportunity to complete by March 21, 2019). This indicates the missing data is not missing completely at random, so the analysis must rely on the missing at random assumption in this dataset, i.e., missingness could depend on covariates and/or earlier observed post-baseline CDRSBs**.** In study 301 also, the proportion of region 1 decreased steadily over the course of the study and was decreased by more than 15% and the proportion mild increased by 10% and there were 8% more placebo than high in the youngest group and 9% more high than placebo in the middle age group

for those randomized after Sep 21, 2017. Similarly, in the PV4 subset (those who consented to protocol version 4 [dated March 2017] by Week 16 of their participation), 97% of those with the opportunity to complete had baseline disease stage Prodromal/MCI, whereas among those without the opportunity to complete 74% were Prodromal and 26% were Mild AD. This is illustrated in Figure 1 which shows Study 302 enrollment probability of Mild (rather than Prodromal) AD at baseline over the duration of enrollment. The red and blue curves are based on a generalized additive model for a binomial response, i.e., for the baseline disease stage with 0 representing Prodromal/MCI and 1 representing Mild AD baseline disease stage, to fit the local trend for probability of Mild baseline disease stage enrollment as a function of calendar time for pre-PV4 (blue) and post-PV4 (red) separately. The curves indicate that the probability of Mild baseline disease stage started high then decreased to near zero and began to increase again slightly as PV4 patients were enrolled. This indicates that PV4 consented patients with the opportunity to complete were more Prodromal than pre-PV4 and that those PV4 without the opportunity to complete were typically more Mild than those PV4 with the opportunity to complete. The opportunity to complete population analysis seems the most appropriate since it relies least on the model being correctly specified and considering the various suggestions from the data that the model may have failed to include various important interaction effects (Country*VISIT [.0030], Baseline Disease Stage*Visit [<0.0001], Baseline AD meds*VISIT [0.0006], Weight*VISIT [0.0118], APOE*VISIT [0.0694]) which could cause the ITT analysis to be biased given the large amount of missing data (>40% overall with 70% missing post-PV4).

.

Figure 1. Probability of Randomization of Mild (1) vs. Prodromal (0) Baseline Disease Stage as Study 302 progressed

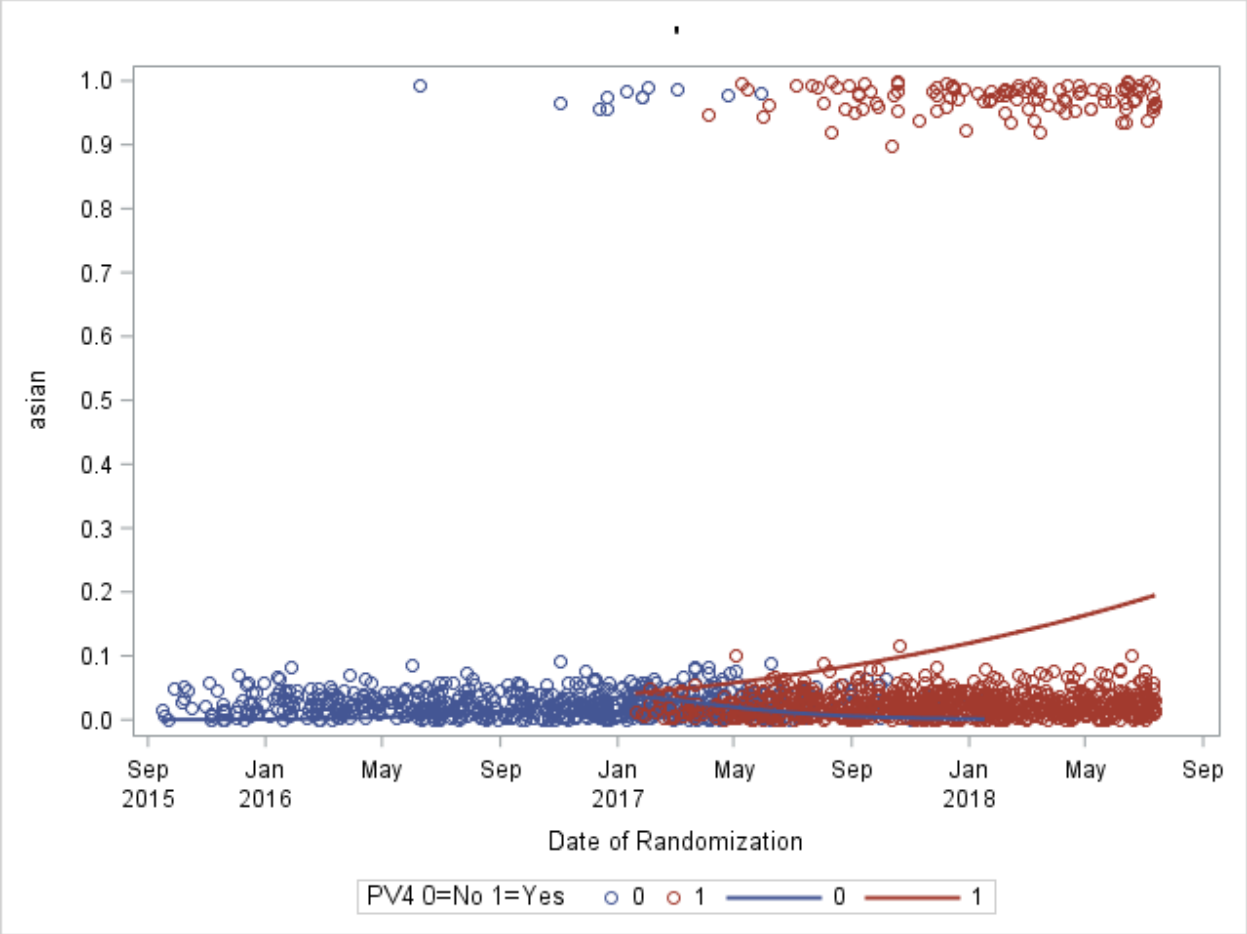


Note: The red and blue curves are based on generalized additive models for a binomial response, i.e., the baseline disease stage with 0 representing Prodromal and 1 representing Mild AD baseline disease stage for pre-PV4 (blue) and post-PV4 (red) separately, These models help to reveal the local trend for probability of Mild baseline disease stage enrollment as a function of calendar time.

Figure 2 shows that the probability of a randomized patient being Asian increased post-PV4 (red) in Study 302.The blue symbols indicate pre-PV4 patients and the red symbols indicate post-PV4 patients. In the figure shown, y=0-values represent non-Asian patients and y=1 values represent Asian (a small amount of noise was added to avoid obscuring many coincident points). The curves show the trend in the probability over time as estimated by a generalized additive binomial model (smooth curve) for the probability of a randomized patient being Asian.

Figure 2. Probability of Asian Enrollment over Time in Study 302

0= Non-Asian ; 1=Asian Blue=Pre-PV4 Red=Post-PV4



Post-Pv4 patients compared to pre-PV4 patients also had some of these baseline characteristic differences as shown in Table 8.

Table 8. Study 302 Demographic Characteristics by pre-PV4 and post-PV4

		PV4		All
		Pre-PV4	Post-PV4	
Baseline Alz Dis Med Use Flag				
No	**N**	323	467	790
	Percent	44.74	50.98	48.23
Yes	**N**	399	449	848
	Percent	55.26	49.02	51.77
Baseline Alzheimer Disease Stage		129	173	302
MILD	**N**			
	Percent	17.87	18.89	18.44
PRODROMAL (MCI)	**N**	593	743	1336
	Percent	82.13	81.11	81.56
Geographic Region 1				
Asia	**N**	12	109	121
	Percent	1.66	11.90	7.39
Europe/Canada/Australia	**N**	383	482	865
	Percent	53.05	52.62	52.81
United States	**N**	327	325	652
	Percent	45.29	35.48	39.80
All	**N**	722	916	1638

In addition to the enrollment changes in region over time there was in fact a significant three way interaction between country, treatment and visit; comparing these additional model terms to the primary analysis model, the three way interaction was highly significant F=1.63 (num df=106, den df=3024) p<.0001 and results for related terms tested individually were:

EFFECT	F statistic	p-value
COUNTRY	4.45	<.0001
COUNTRY*TR01PG1N	0.78	0.7644
COUNTRY*AVISITN	2.00	0.0028
COUNTRY*AVISITN*TR01PGN	1.55	0.0095 .

The primary analysis of study 302 contained 1581 patients and 3716 observations. About 14% of randomized patients in study 302 started concomitant AD medications post-baseline (the highest use was in the high dose 15.2% [17.8% post -PV4]). The overall rate of starting increased slightly after the protocol version 4 implementation 11.4% before to 16.0% after, including 12 before to 17.8% after for the high dose [study 301 overall (10.0% pre to 13.3% post)]. For the

final BLA data the proportion starting concomitant AD medications post-baseline increased relative to the June 2019 dataset from 10.8% (June 2019) to 14.1% (BLA). Sensitivity analyses addressing post-baseline starting of concomitant AD medications involved the following number of patients and results (shown in Table 9).

Table 9. Study 302 Week 78 CDRSB Analyses Exploring Exclusion of Data after post-Baseline starting of AD medications

Handling	Patients	Records	High vs. Placebo LS Mean Difference	Std. Error of Difference	p-value
Censoring impacted Data	N=1524	3466	-0.410	0.158	0.0098
Censoring Impacted Patients	N=1358	3206	-0.431	0.161	0.0074

In study 302, the estimated high dose effect was smaller in those who started concomitant AD medications (-0.118 [S.E.=0.501] vs. -0.431[S.E.=0.161] in those who did not). In study 301 the subgroup that started concomitant AD medications was in the right direction, but the subgroup that did not was in the wrong direction for the high dose compared to placebo (-0.272 [S.E. =0.562.] and +0.088 [S.E.= 0.153], respectively).

The sponsor makes an argument about outliers being influential more in study 301 than 302. However, a robust regression, which is by design less affected by outliers (least trimmed squares /M estimation), of study 301 Week 78 data shows no effect of the high dose on CDRSB (Placebo-High = 0.0265, S.E. = 0.1247, p=0.8315). The low dose is also still numerically better than the high dose in the robust analysis of study 301 (-0.0628, S.E. = 0.1249, p=0.6153). The Opportunity to Complete Subset as defined by the Sponsor consists of those randomized patients who were randomized early enough in the study timeline in order to have had the Opportunity to have a Week 78 assessment before March 20, 2019. Note that missing data is fairly limited in the Opportunity to Complete subset (90.5% complete in 301 and 91.9% complete in 302) so that analyzing just the Week 78 Visit in this subset seems not unreasonable for the exploratory robust regression, especially since there is no indication of an earlier effect (Week 26 or Week 50) in the primary MMRM analysis.

Figure 3 shows histograms of Week 78 changes from baseline in CDRSB for placebo and the high dose with the bars side by side at each level of change in the Opportunity to Complete Population. One can determine from Figure 3 the likelihood of observing specific changes from baseline in CDRSB at Week 78 based on the totality of phase 3 trial evidence among those randomized patients who had the opportunity to complete 78 Weeks.

Figure 3. **CDRSB changes at Week 78 in those with opportunity to complete (studies 301 and 302 pooled)**

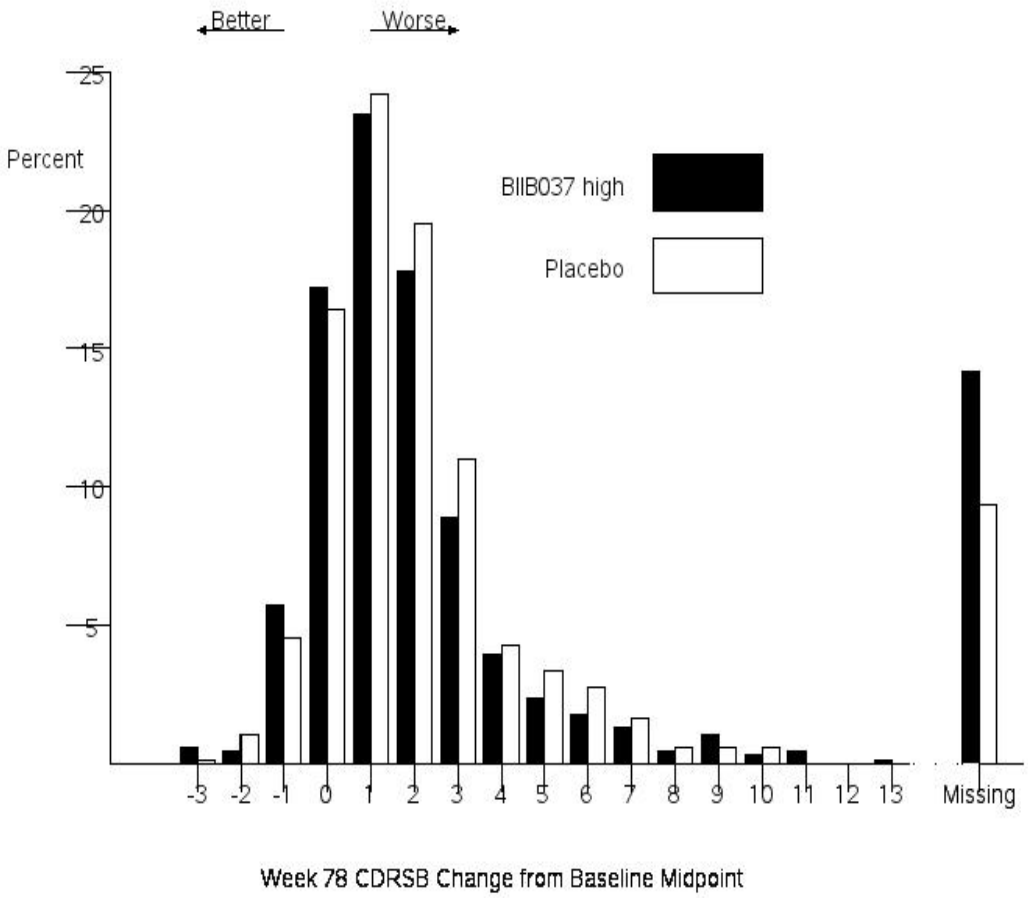
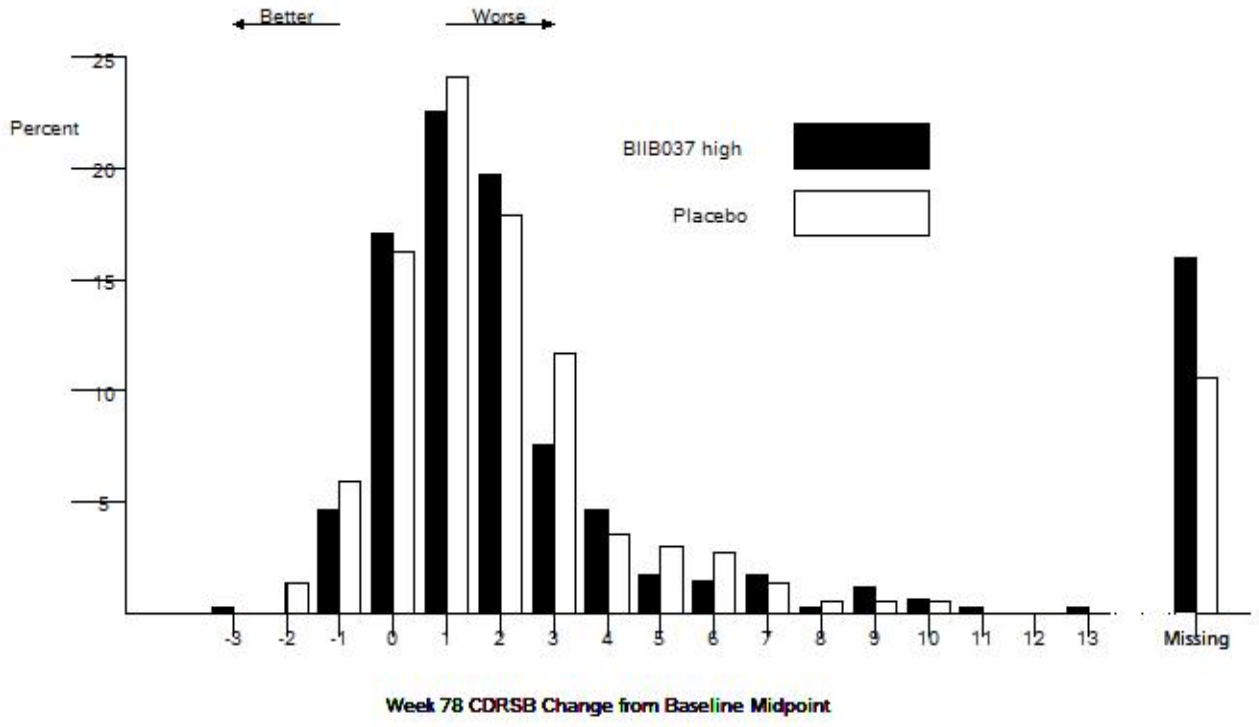


Figure 4 shows histograms for CDRSB Changes at Week 78 in Study 301 for the high dose and placebo groups. One can determine from Figure 4 the likelihood of observing specific changes from baseline in CDRSB at Week 78 based on the totality of phase 3 trial evidence among those randomized patients who had the opportunity to complete 78 Weeks.

Figure 4. Study 301 CDRSB Changes at Week 78 (excluding missing data but deaths coded as missing)



The sponsor prespecified two responder thresholds for secondary analyses of CDRSB, 1.5 and 0.5 thresholds

In 302 proportions of responders for placebo vs. high dose at Week 78 were as follows.

ITT population: missing treated as non-responder

CDRSB <=0.5 18.8% vs. 25.7% p=.0029

CDRSB<= 1.5 32.2% vs 39.1% p=.0099

Opportunity to Complete Population:

CDRSB <=0.5 32.8% vs. 40.7% p=.0135

CDRSB<= 1.5 56.2% vs 62.1% p=.0550

In 301 these were as follows.

ITT population: missing treated as non-responder

CDRSB <=0.5 25.7% vs. 20.2% p=.7334

CDRSB<= 1.5 39.9% vs 36.8% p=.1621

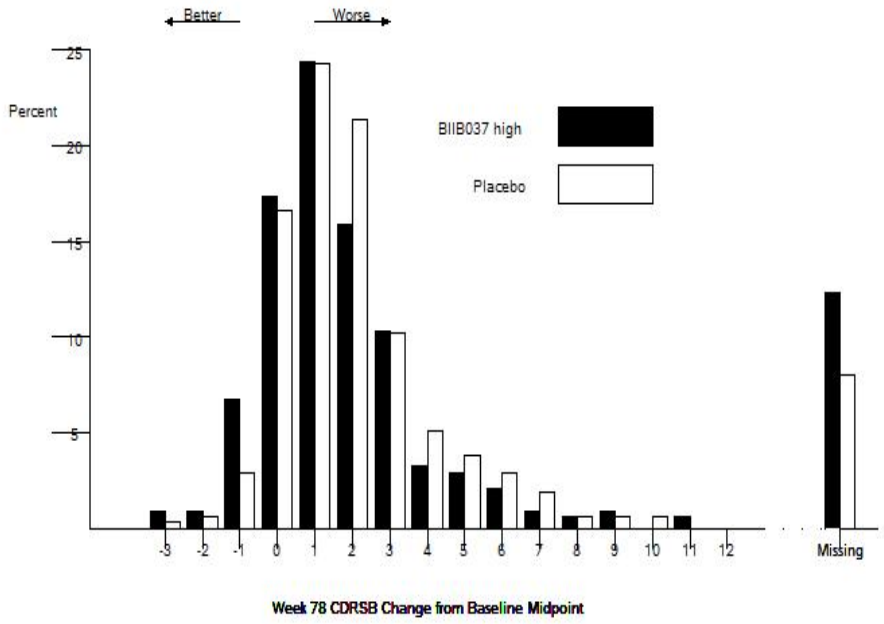
Opportunity to Complete Population:

CDRSB <=0.5 37.7% vs. 32.2% p=.3332

CDRSB<= 1.5 58.7.2% vs 59.1% p=.5986

In Figure 5 which shows the observed distribution of Week 78 CDRSB changes for placebo and high dose groups one can see after referring back to Figure 4 for Study 301 that Study 302 only had one fewer worse change observed for the high dose, i.e., a 13 point worsening, (and 301 had one 11 point worsening while study 302 had two 11 point worsenings).

Figure 5. Study 302 CDRSB changes at Week 78 (excluding missing data due to futility stopping)



The sponsor also asserts that there is no bias due to ARIA because excluding data after ARIA doesn't markedly change the primary result. However, one can't conclusively rule out an impact of those experiencing ARIA on the result because it requires making a comparison based on differential exclusions between the randomized groups (drug patients and/or censoring of drug arm data) and the resultant groups without ARIA to be compared are no longer as randomized and/or have differential follow-up and selection bias due to conditioning on this post-randomization event.

Figure 6 shows from left to right a comparison of pre-to post PV4 CDRSB profiles LSmeans by group: placebo (left), low dose (middle), and high dose (right). It shows a dramatic worsening of placebo post-PV4 (red) relative to pre-PV4 (blue). One can see in Figure 6 that placebo was dramatically worse in the APOE+ stratum post-PV4 for CDRSB as compared to pre-PV4, while the the low dose was slightly better and the high dose was slightly worse from pre-PV4 to post-PV4 . The low dose being numerically better than high is the opposite of what would be expected, since only the high dose got a dose increase for post-PV4. Since placebo worsened significantly compared to pre-PV4 and the high dose is numerically worse than low post-PV4 it suggests that placebo worsening after PV4 could be the driver of the overall result.

Figure 6. Study 302 Placebo, Low Dose, and High dose CDRSB profiles by Pre-PV4 and-Post PV4 status (LSMean +/- 1SE) in APOE carriers stratum

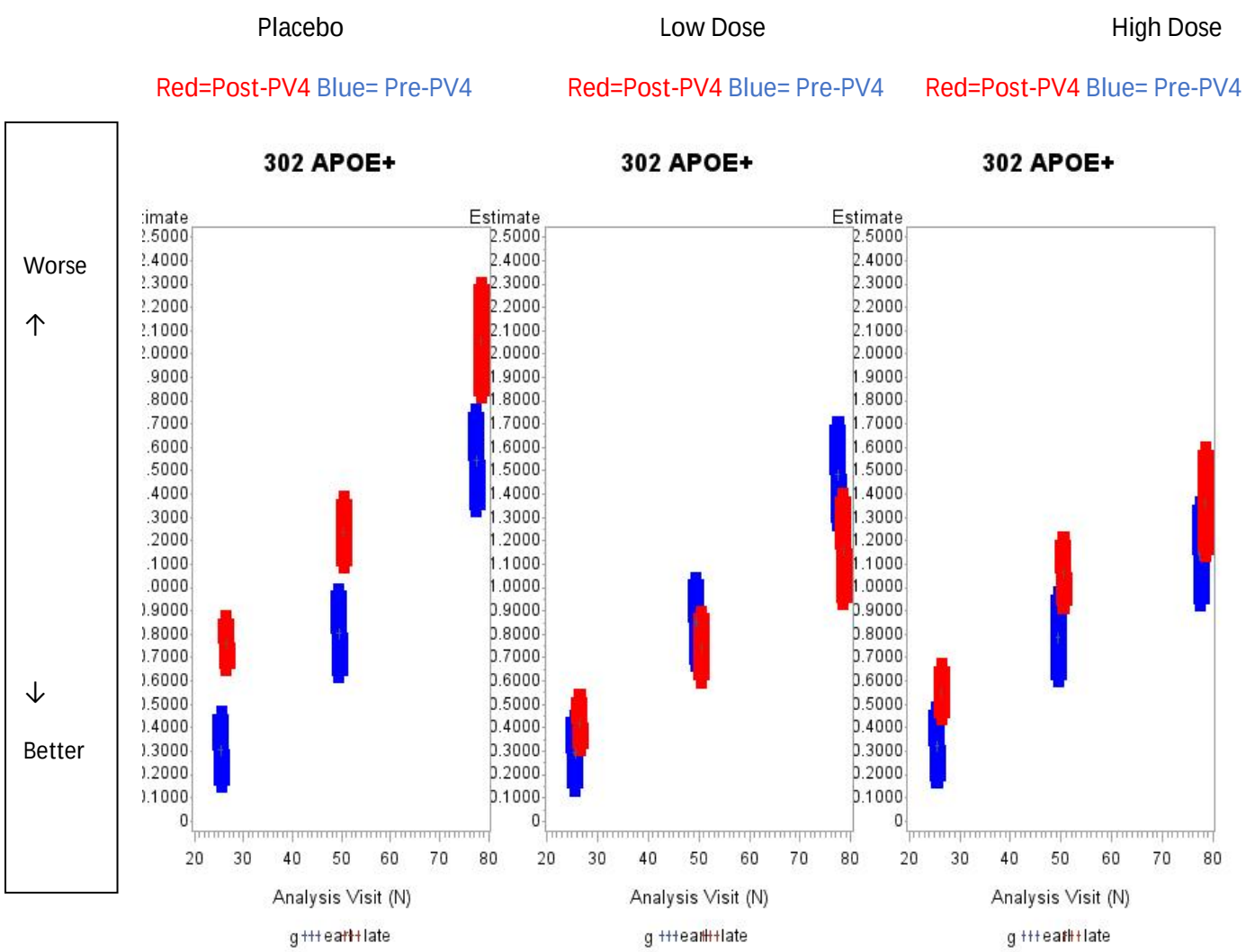


Figure 7 shows the corresponding CDRSB profiles for the APOE non-carrier stratum. The high dose is numerically worse than placebo at Week 78 in the protocol amendment 4 subgroup in the non-carrier stratum (compare red on the left at Week 78 with red on the right at Week 78). Furthermore, in the APOE non-carriers the high dose was only slightly better than placebo in pre-PV4 patients (and the figure shows that despite the titration to 10 mg/kg the high was essentially no better than the low dose pre-PV4).

Figure 7. Study 302 Placebo, Low Dose, and High Dose CDRSB Profiles by Pre-PV4 and Post-PV4 in APOE non-carriers stratum

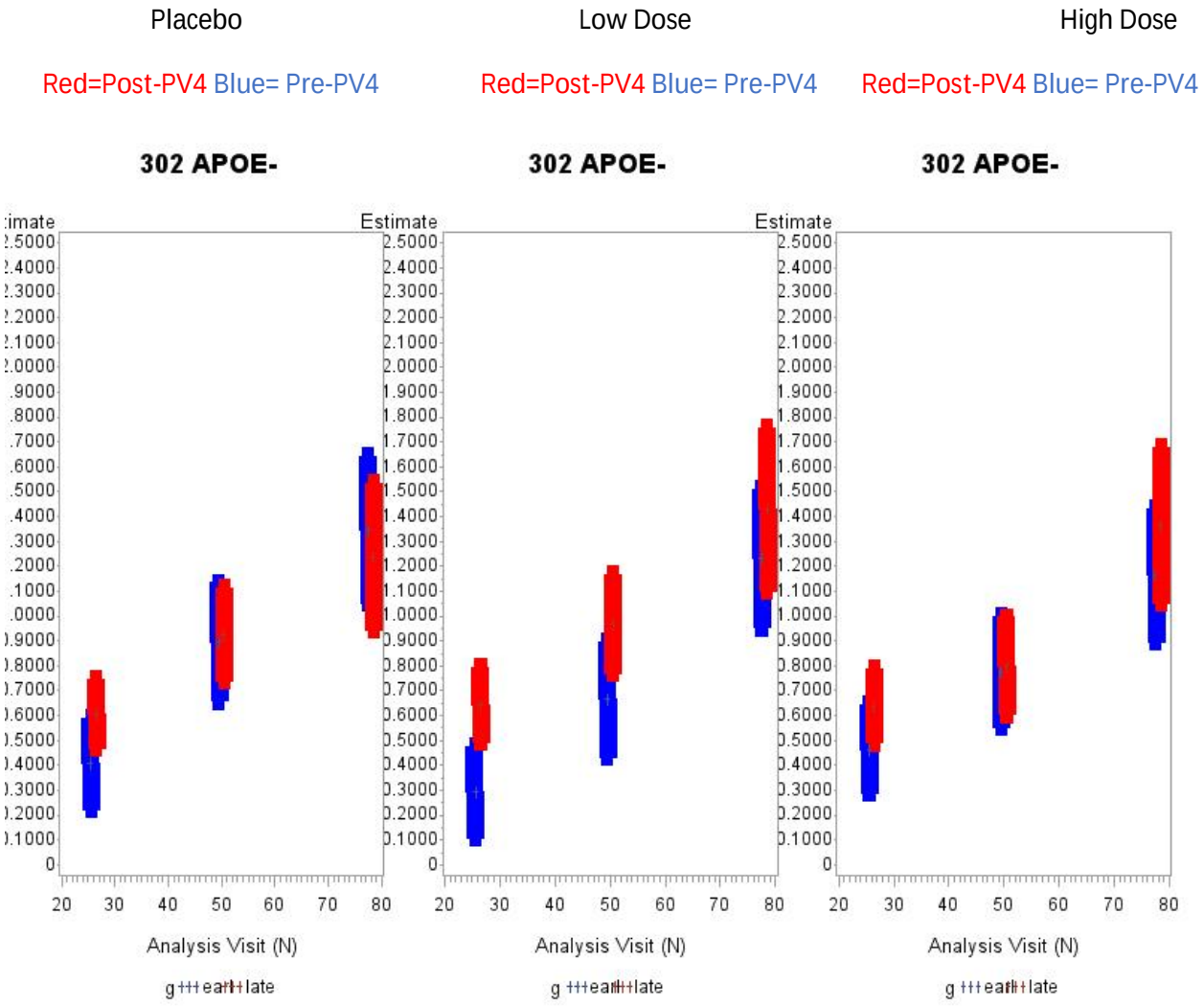


Figure 8 shows the local trend (adjusted for primary model covariates other than Visit since this is all Week 78 data) in study time in LS Mean CDRSB Change at Week 78 within a time window focused around the PV4 implementation. Blue is non-PV4 placebo, Red is for non-PV4 High Dose, Green is for PV4 placebo and brown is for PV4 high dose. The number of 10 mg/kg doses are shown just above and to the right of the plot symbol. One can see that there were some relatively poor Week 78 changes among those PV4 high dose with the full possible 10 mg/kg dosing. The figure suggests that there was a worsening trend in PV4 placebo as well as to a lesser degree PV4 high dose.

Figure 8. Local Trend in CDRSB Week 78 Changes over Study Duration and by PV4

0=Pre-PV4 Placebo ; 1=Pre-PV4 High Dose; 2=Post-PV4 Placebo; 3= Post-PV4 High Dose

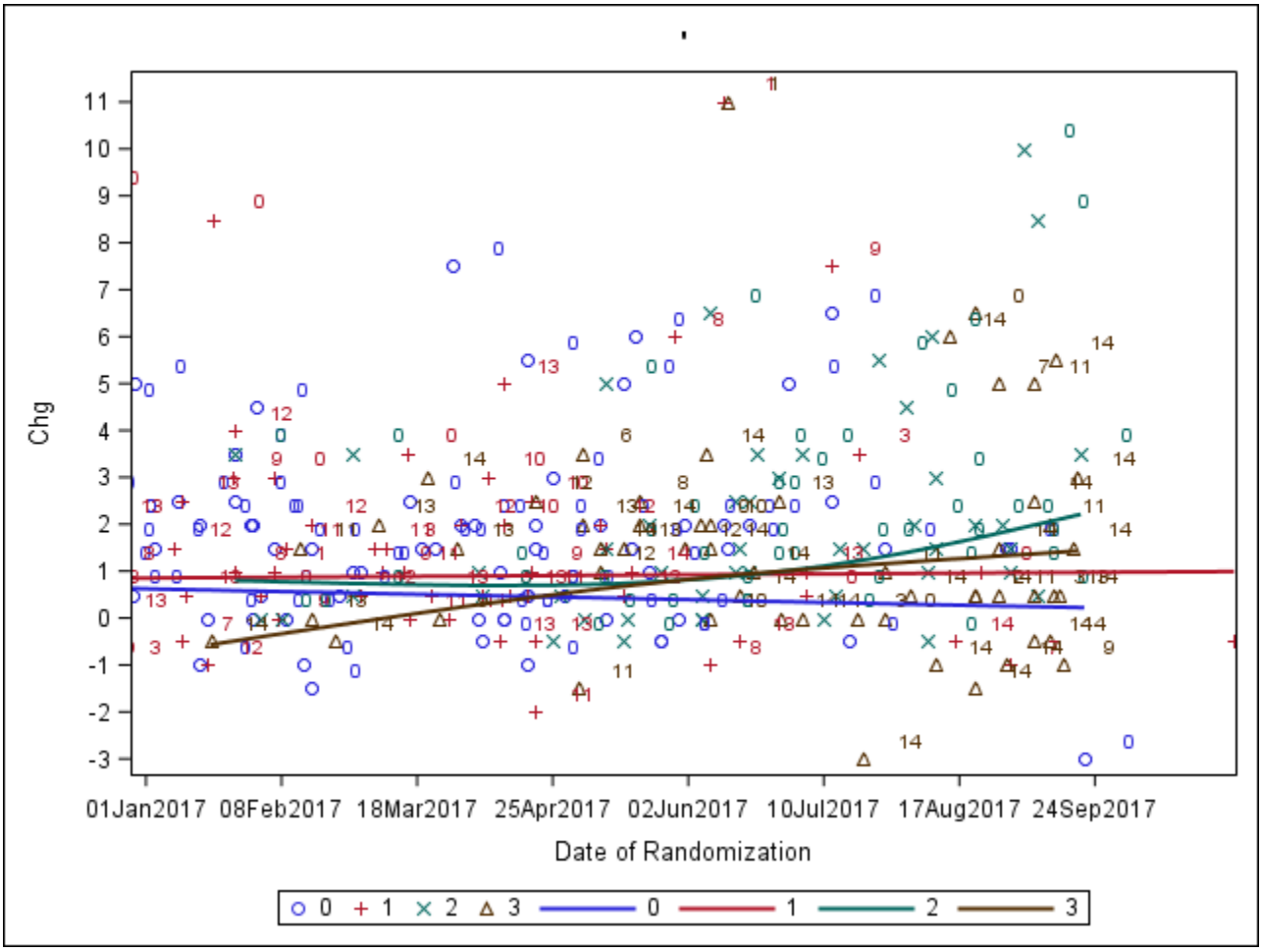


Figure 9 shows the same trends over the full course of the study rather than more focused on the PV4 implementation time above.

Figure 9. Local Trend in CDRSB Week 78 Changes over Study Duration and by PV4

0=Pre-PV4 Placebo ; 1=Pre-PV4 High Dose; 2=Post-PV4 Placebo; 3= Post-PV4 High Dose

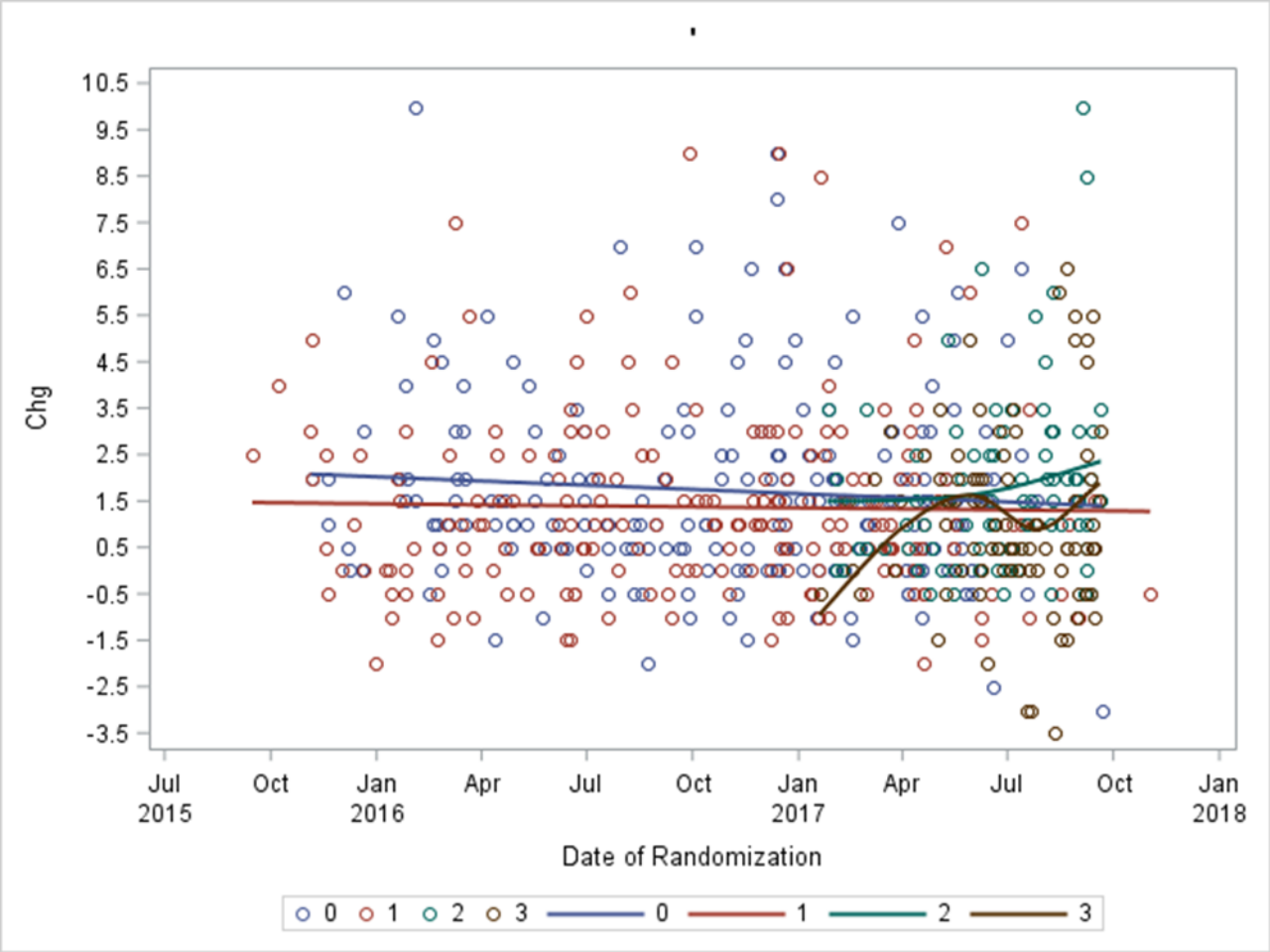
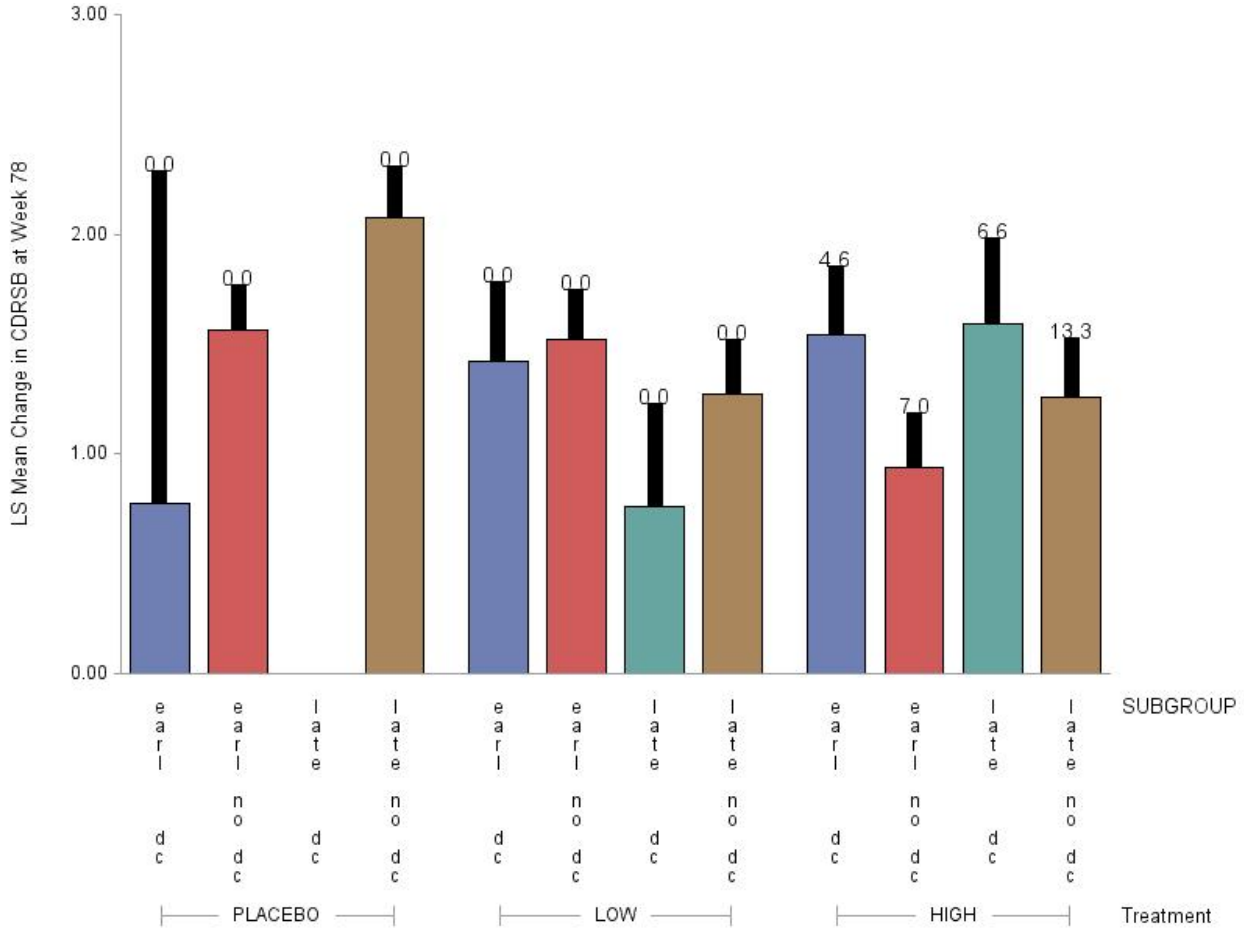


Figure 10 shows subgroup analyses by dose group and dose titration modification due to ARIA adverse events subgroups comparing pre-PV4 and post-PV4 in APOE carriers. The black bar extensions represent the standard errors of the bar heights. The statistic over the bar indicates the average number of 10 mg/kg doses which is obviously zero for the low and placebo doses (ranging from 4.6 to 13.3 for high dose). DC stands for dose titration change or at least modification due to an adverse event of ARIA. Those classified as early by not meeting the sponsor's definition of PV4, which requires consent by Week 16 could still possibly have consented to PV4 later than Week 16 (depending on enrollment time). We can see that the late

dose modified due to ARIA subset for the low dose had a better mean than the late non-dose modified/normal titration subset of the low dose (in the June 2019 dataset the late high dose showed the same surprising trend, i.e., dose modified subgroup for high numerically better than high normal titration subgroup across these subgroups) and it is interesting to note that the late high dose subgroups are no better than the corresponding low dose groups. Overall, in PV4 APOE+ the low dose LS mean at week 78 is 1.16 and the high dose LS mean is 1.34. Worsening of placebo from early to late alone, can potentially explain the high doses improved outcome relative to placebo, i.e., the high dose is not improved relative to the corresponding low dose and placebo worsened. In study 302, the late non-dose reduced due to ARIA subgroup is also numerically worse than the Early High Dose non-dose-reduced subgroup (red) that also had significantly fewer 10 mg/kg doses by Week 78 and essentially the same LSmean as the corresponding late low dose (middle brown bar). Overall in study 302 APOE+ post-PV4, regardless of dose titration changes or lack thereof, at Week 78 the high dose LS mean change from baseline in CDRSB was worse than the low dose calling into question the supposed importance of 10 mg/kg doses (and High essentially the same as Low in study 301 post-PV4 APOE carriers). What did change dramatically in study 302 was the placebo response was considerably worse post-PV4 (pre: 1.54 [S.E.= .21] vs. post: 2.06 [S.E.=.24] this would be a nominally significant change; compare leftmost brown and blue and red bars [slightly less worse 1.71 and 2.04 if using all data with a PV4 interaction instead of separate subgroup analysis]). There also appears to be very little difference between the low (middle brown) and high dose (right brown) LS means post-PV4 after the dose of the high dose had been allowed to increase from 6 to 10 mg/kg.

The sample sizes shown below the figure are admittedly small but in the June 2019 data the late high dose unmodified subgroup had the surprising result that the high dose subgroup with dose titration change due to ARIA adverse events had a numerically better Week 78 LS mean CDRSB change than the complement unmodified subgroup that had more 10 mg/kg doses (a numerically worse LSmean than the late high dose titration slowed subgroup N=37 1.28 [S.E.=0.27] vs. N=15 1.13 [S.E.=0.41]). However, in the final data there were 4 high dose that had been designated pre-PV4 in the June 2019 data that had become post-PV4 and this trend was reversed as seen in Figure 10 N=38 1.25 [S.E.=0.27] vs. N=13 1.59 [S.E.=0.39] likely due to the week 78 CDRSB being changed from 1.0 to 3.5 for one dose reduced or slowed high dose patient and from -0.5 to -1.0 for one patient in the non-dose titration modified subgroup. Only the final visit, Week 78, CDRSB value changed for each of these two patients and the change was submitted after the general unblinding in April 2019. Still in the final data the dose titration slowed due to ARIA subgroup of the low dose's LSmean at Week 78 is numerically better than the low dose titration unmodified as well as both the high dose modified and high dose unmodified titration subgroup LSmeans.

Figure 10. Study 302 Week 78 LS Mean CDRSB Change by ARIA Dose Modification Status and pre/post PV4 in APOE+



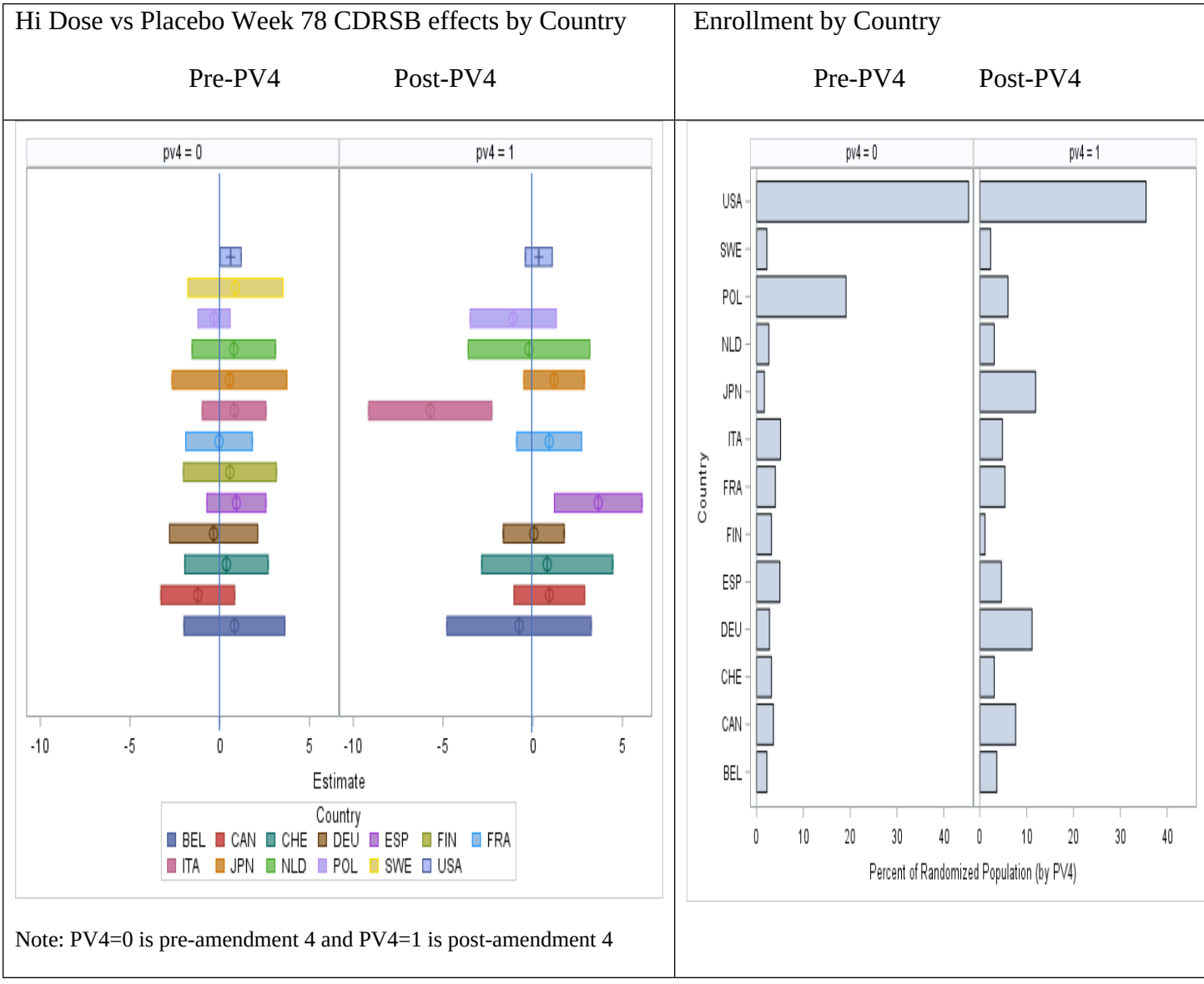
- N=2 N=151 N=48 N=36 N=106 N=13 N=43 N=50 N=89 N=18 N=38
- Note: earl=Pre-PV4 late=Post-PV4; dc= dose titration slowing or reduction due to ARIA

Overall, not subsetting by PV4, the country by Visit by treatment group interaction was significant, 0.0095 (had been p=0.01 in the June 2019 data), possibly suggesting that CDRSB profiles over time and treatment group differences varied significantly by Country, especially considering the low power for interaction tests and that the test of all interaction terms involving country, beyond the country main effect, being zero has a p-value of 0.0024, suggesting that collectively the interaction terms are highly significant. Exploratory exclusion of Japan, which had significant data irregularities of various sorts reaching a level designated as atypical (as identified by Dr. XiaoFeng Wang, a statistical analyst in the Analytics and Informatics staff of

the Office of Biostatistics using Cluepoints software), there are 1462 (93%) remaining patients and the high dose vs. placebo difference on CDRSB at Week 78 in 302 is -.33 (.16) with p-value increasing to 0.040. Exploratory exclusion of Spain (1504 or 95% remaining patients) results in a high difference of -30 (.159) with p-value of 0.0599.

Figure 11 shows the distribution of enrollment by Country pre-PV4 and post-PV4 on the right, as well as the differences in estimated high dose Week 78 effects by Country pre-PV4 and post-PV4 on the left. The post-PV4 subset also had a significant country by visit by treatment interaction p=0.0178, suggesting that treatment differences across visits were not consistent across countries in the post-PV4 subset.

Figure 11. Differences in estimated Week 78 CDRSB High Dose effects by Country in Study 302



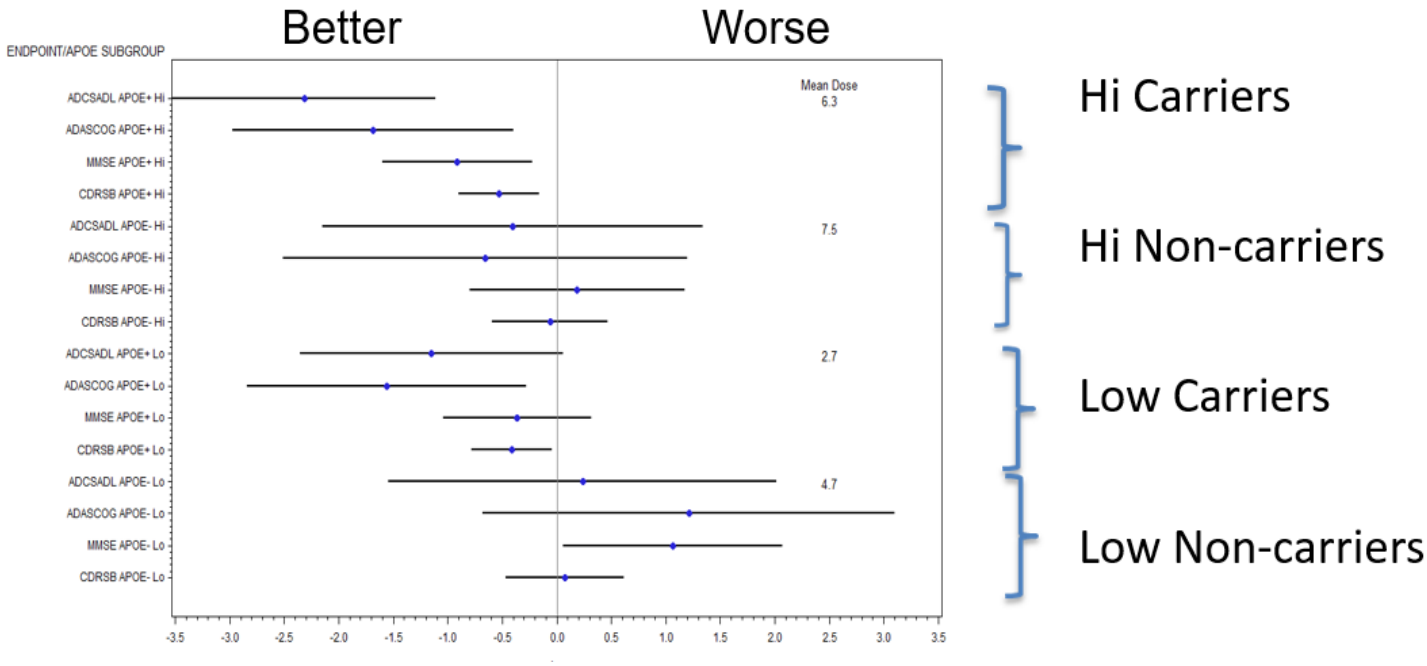
Note: PV4=0 is pre-amendment 4 and PV4=1 is post-amendment 4

Considering the forest plot below there is very likely less happening in APOE non-carriers in **study 302** although they got higher dosing from the beginning of study conduct..

There are eight possible comparisons of Week 78 treatment effect estimates between APOE+ to APOE-, across the four key endpoints and two doses. The forest plot (Figure 12), cycles through the four endpoints from top to bottom first for APOE + high dose, then for APOE- high dose, then for APOE+ low dose, and, finally, the four endpoints for APOE- low dose. One can see that for each horizontal line the point estimate at the center for the opposite APOE subgroup for the same endpoint and the same dose (four bars below) is lower. In eight out of eight cases, APOE+ is better. Note that the signs of MMSE and ADCS-ADL-MCI were changed, so that for all four endpoints a positive difference would favor the drug. All four endpoints for APOE- non-carrier low dose are also in the wrong direction compared to placebo numerically.

In fact, there is a statistically significant interaction for the first key secondary endpoint MMSE between APOE and treatment groups, p=0.0096, i.e., **the treatment effect(s) is not the same, a nominally significant difference, across APOE subgroups**.

Figure 12. Study 302 Key Endpoints APOE subgroup treatment effects estimates by Dose



For all of the four key endpoints the APOE- estimated effect is lower than APOE+ for both low and high doses (see also Table 10). This is despite the fact that APOE- high dose group had 10mg/kg as maximum dose both pre-PV4 and post-PV4 (thus higher average exposure than APOE+). On the first key secondary endpoint (MMSE) the treatment by APOE interaction test has a p-value of 0.0096.

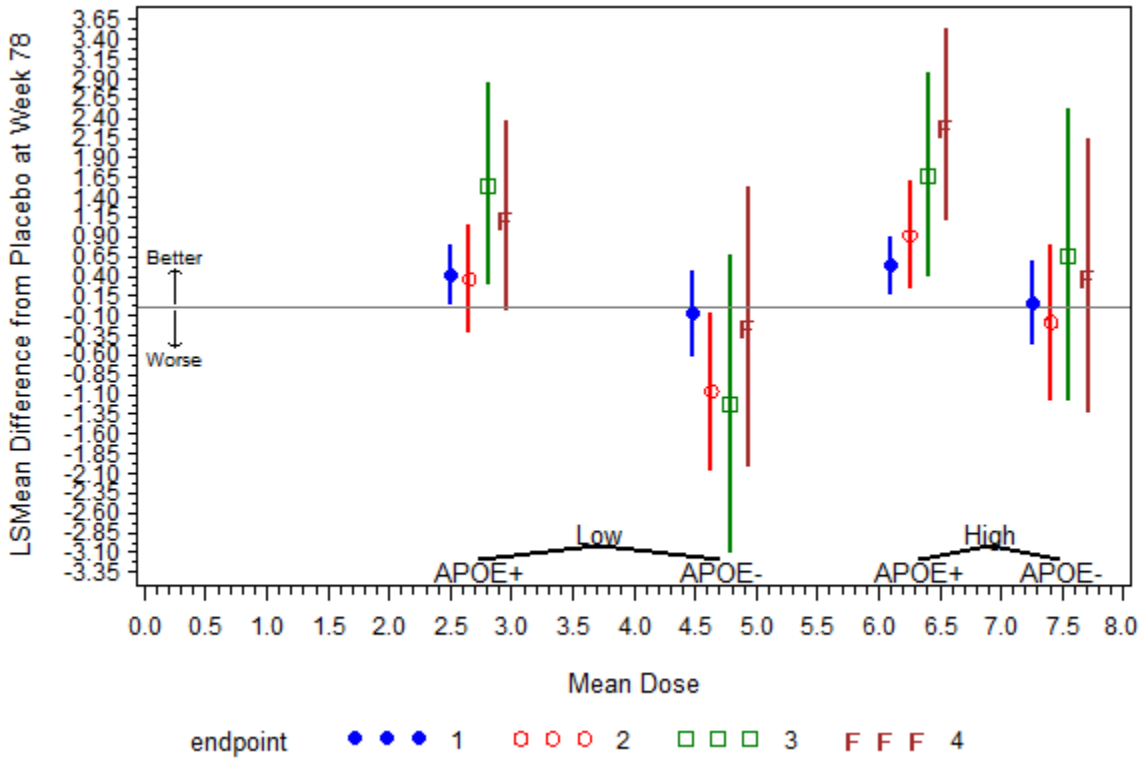
A test for a difference in MMSE treatment effect at Week 78 in either low or high between APOE+ and APOE- has a p-value of 0.0454. If the doses are combined the p-value for consistency of effect across APOE is 0.0187, so that again consistency would be rejected.

Table 10. Study 302 Estimated Treatment Differences at Week 78 by APOE for Primary and Key Secondary Endpoints

Endpoint/Dose	APOE+ Diff vs Pl LSMean (Std.Error)	p-value	APOE- Diff vs Pl LSMean (Std.Error)	p-value	APOE- - APOE+ diff LSMean (Std.Error)	p-value
cdrsb high	-0.53(0.19)	0.0048	-0.06(0.27)	0.8232	0.47(0.33)	0.1531
cdrsb low	-0.41(0.19)	0.0283	0.07(0.28)	0.7878	0.49(0.34)	0.1458
mmse high	-0.91(0.35)	0.0092	0.19(0.50)	0.7118	1.10(0.61)	0.0734
mmse low	-0.36(0.35)	0.2985	1.06(0.51)	0.0386	1.42(0.62)	0.0217
adascog high	-1.69(0.66)	0.0103	-0.66(0.95)	0.4869	1.03(1.15)	0.3714
adascog low	-1.56(0.65)	0.0172	1.21(0.96)	0.2087	2.77(1.17)	0.0175
adcsadl high	-2.32(0.62)	0.0002	-0.40(0.89)	0.6491	1.91(1.08)	0.0772
adcsadl low	-1.15(0.62)	0.0619	0.24(0.91)	0.7942	1.39(1.10)	0.2063

A test for a difference in estimated high dose treatment effect at Week 78 between APOE+ and APOE- has a p-value of 0.073 for the first key secondary MMSE and 0.077 for the third key secondary endpoint ADCS-ADL-MCI. For the low dose, a test for a difference in estimated treatment effect at Week 78 between APOE+ and APOE- has a p-value of 0.022 for MMSE and 0.018 for ADASCOG13. All of these trends suggest bigger effects in APOE+ as compared to APOE-. This is further illustrated in Figure 13 which plots LS Mean difference from Placebo versus Mean Dose. We can see that surprisingly despite the higher dose for non-carriers than carriers, non-carriers have little treatment effect (Placebo-Drug, so that higher is better in the figure) and, in fact, all key endpoint differences from placebo are numerically worse than placebo for the low dose. In the absence of an APOE by Treatment interaction the four vertical lines with the same symbol and color (one set for each of the key endpoints) with the mean identified at the center should be consistently monotonically increasing from left to right, with increasing dose, but this is not the case, for any of the key endpoints.

Figure 13 Interaction between CDRSB and APOE consistent across Dose groups

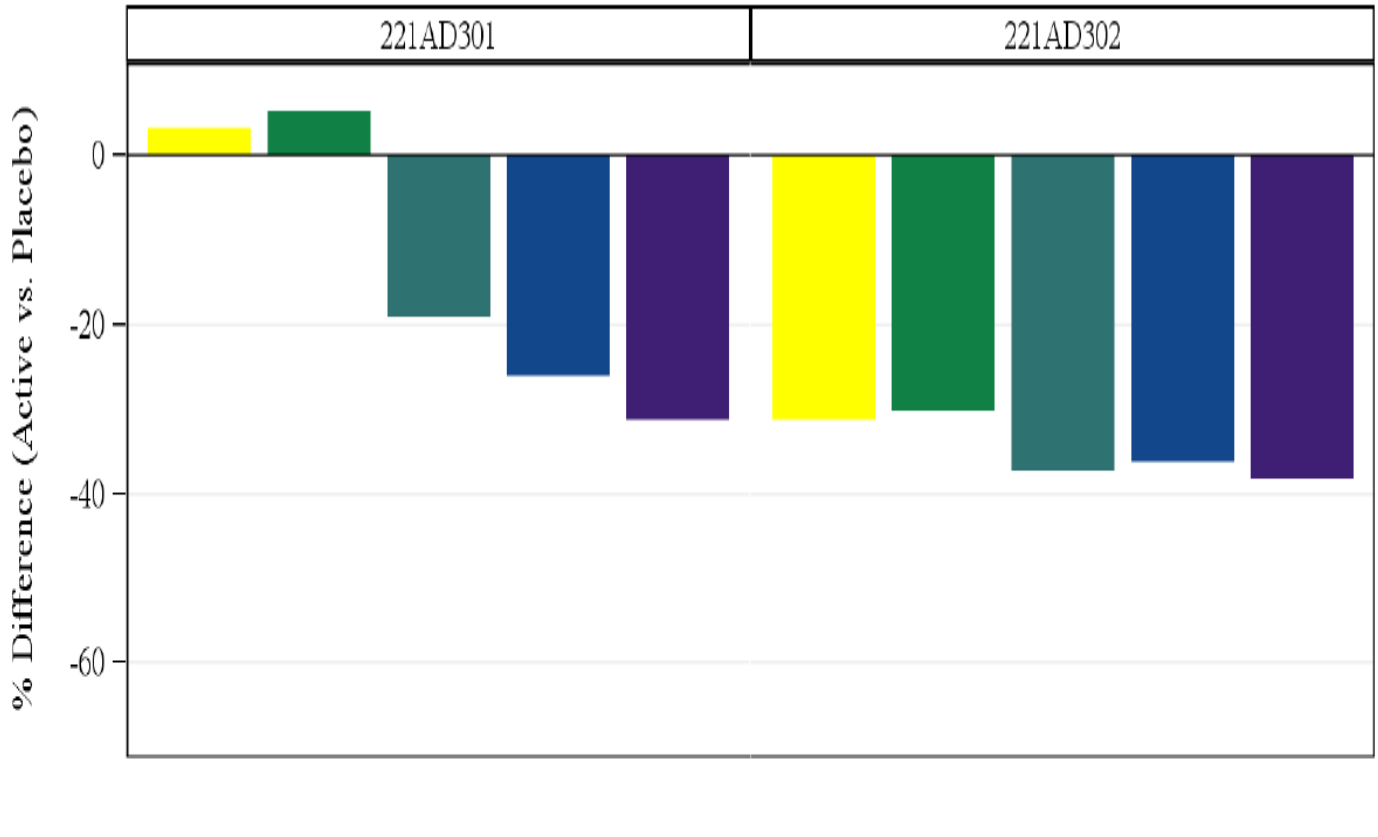


Note: endpoint 1=CDRSB 2=MMSE 3=ADCS-Cog13 4=ADCS-ADL-MCI

The sponsor tried to use a post-hoc propensity score matching analysis to suggest that there was a trend in study 301 for the high dose as a function of the number of 10 mg/kg doses(Figure 14). The sponsor did not include standard errors of the bars, so significance of differences are

difficult to judge. However, one can notice that study 302 shows no real difference between >=6,>=8, >=10,>=12, or =14, which would suggest that the number of 10 mg/kg doses doesn't matter in study 302. On the other hand, the sponsor seems to suggest that in 301 the number of doses matters (an apparent linear trend in the bars). However, the linear trend in % difference in 301 may well be attributable to the simultaneous linear trend in placebo responses across the categories (>=8 to =14 : varying from 1.36 at the second lowest category [>=8] to 1.58 at the highest [=14], i.e., the worst placebo response coinciding with the highest dose category). Furthermore, overall, the figure suggests again that the studies are not consistent (with respect to the importance of the number of doses in this case). The sponsor provided no accompanying assessment of how well the resulting groups were matched in the graph and it could be poor since the number of like patients decreases with the number of matching factors and, for example, enrollment window of every 200 subjects was a matching factor. Also, this reviewer has shown that the CDRSB outcomes in the non-US countries were not sufficiently similar to be pooled but this analysis pooled them. Therefore, this analysis does not incorporate these important regional differences into the matching and therefore the analysis may be confounded with regional differences in estimated Week 78 effect, on top of the unstable placebo response across the categories in 301. There are also differences in estimated Week 78 effect by APOE which this analysis does not properly address. The implications of the figure are also called into question by the fact that the low dose was numerically better than the high dose in Study 301 despite having zero 10 mg/kg doses! In summary this post-hoc subgroup analysis is flawed and such a post-hoc subgroup analysis could never measure up to a prespecified primary analysis supported by randomization, i.e., the overall prespecified final analysis of 301.

Figure 14 Bar plot of CDR Sum of Boxes Adjusted Mean Change from Baseline Percent Difference from Placebo at Week 78 by Number of 10 mg/kg Doses, with Placebo Selected by Propensity Score Matching - ITT Population that have had Opportunity to Complete Week 78 by 20Mar2019: Placebo-controlled Period Excluding Data after 20Mar2019 – Nested Categories



Number of subjects and Adjusted mean at Week 78

Placebo	202	185	157	129	77	220	201	177	144	98
	1.42	1.36	1.49	1.54	1.58	1.56	1.45	1.54	1.38	1.41
BIIB037	202	185	157	129	77	220	201	177	144	98
	1.45	1.43	1.21	1.14	1.08	1.07	1.02	0.97	0.89	0.87

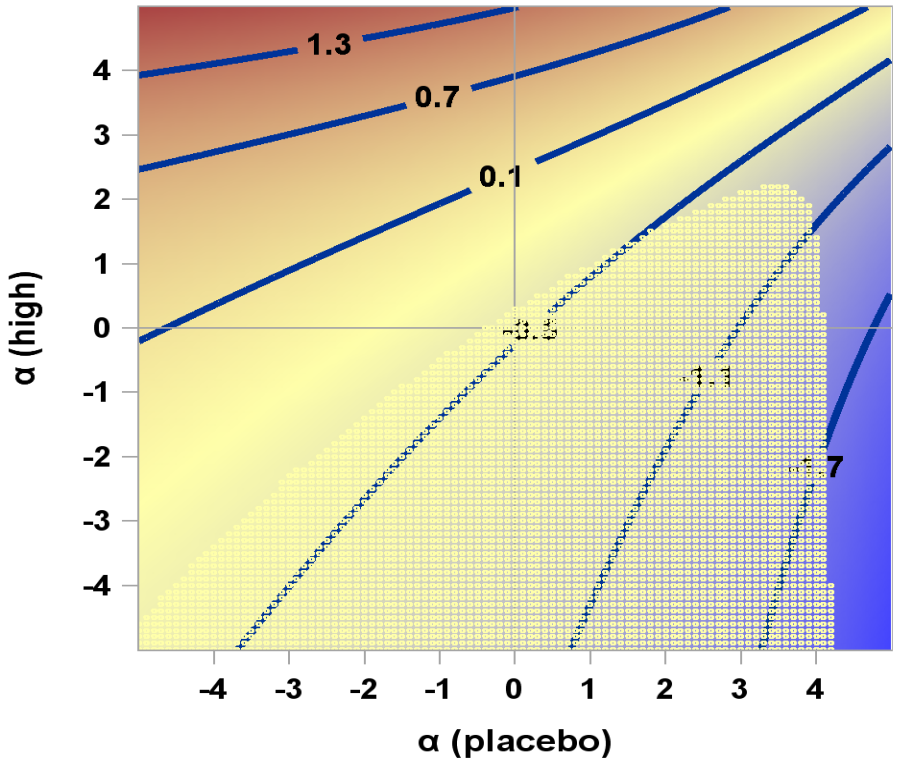
NOTE 1: Covariates in propensity score model include laboratory ApoE status, age, sex, baseline clinical stage, baseline scores of CDR-SB, MMSE, ADAS-Cog 13, ADCS-ADL-MCI, years of education, years since first AD symptom, AD symptomatic medication use at baseline, US/non-US and enrollment window of every 200 subjects. Placebo and treated subjects matched exactly on laboratory ApoE status. Subjects with undetermined laboratory ApoE status are grouped in the randomized ApoE subgroup.

NOTE 2: Results for each threshold were based on an MMRM (mixed model for repeated measures) model, with change from baseline in CDR-SB as dependent variable and with fixed effects of treatment group, categorical visit, treatment-by-visit interaction, baseline CDR-SB, baseline CDR-SB by visit interaction, baseline MMSE, AD symptomatic medication use at baseline, region, and laboratory ApoE status.

Note: The above figure was copied from page 80 of the sponsor's summary of clinical efficacy

Figure 15 shows a tipping point analysis of Study 302, investigating combinations of degrees of informative missingness that could alter the significance (above the shaded grid region) on the primary outcome at Week 78. Alpha ("high") and Alpha (placebo) allow for adjustments to the mean response for dropouts independently for high dose and placebo, i.e., the graph shows the treatment difference when the mean in dropouts is assumed to be the mean in completers for the given group plus an additional amount: + alpha("high") if the given group is high dose or +alpha(placebo) if it is placebo. The figure investigates the potential impact of such possible informative missing data in dropouts. Combinations of the two alpha's above the yellow shaded grid would suggest a loss of significance, e.g., if α(high) is a little greater than 0.3 and α(placebo) = 0 or if α(high)=0 and α(placebo)=-0.43 or any combination of α(high) and α(placebo) above the line α(high) =0.3 +0.7*α(placebo). The same would be true for any combination of α("high") and of α(placebo) in the orange region at the upper left of the figure. Considering the overall outcome of Study 301 it is not unrealistic that high dose dropouts could be worse than placebo.

Figure 15. Study 302 CDRSB Tipping Point Analysis



In Figure 16 since there is no correlation between average achieved dose and outcome when low is plotted with high (not differencing from placebo) it suggests that the sponsor's placebo differenced analysis of correlation between average dose and CDRSB change at Week 78 is driven by changes in placebo outcomes between early and late. Recall that the low dose was numerically better than high by 0.2 points in study 301 at Week 78 although the low dose had a much lower average dose. This lack of correlation also seems to contradict the sponsor's argument about intermediate dosing, i.e., fewer 10 mg/kg doses than 302 undermining the study 301 outcome.

Notice also in the figure below that in 302 the late APOE+ low dose LS Mean which has the plotting symbol **X** is essentially the same as the late APOE + high dose with plotting symbol X.

Figure 16. Assessment of Correlation between Dose Achieved and Week 78 CDRSB across Low and High Doses (not placebo subtracted)

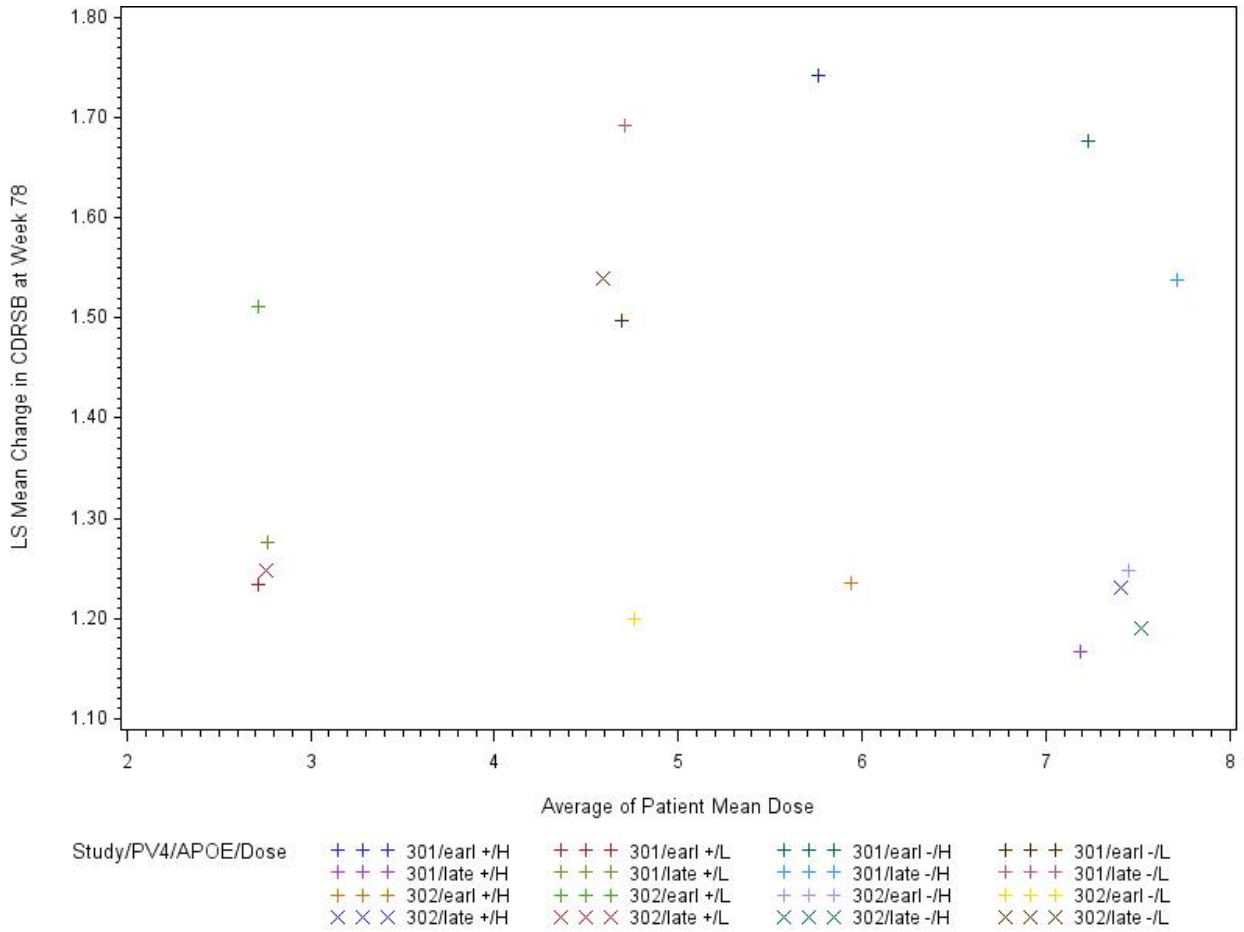
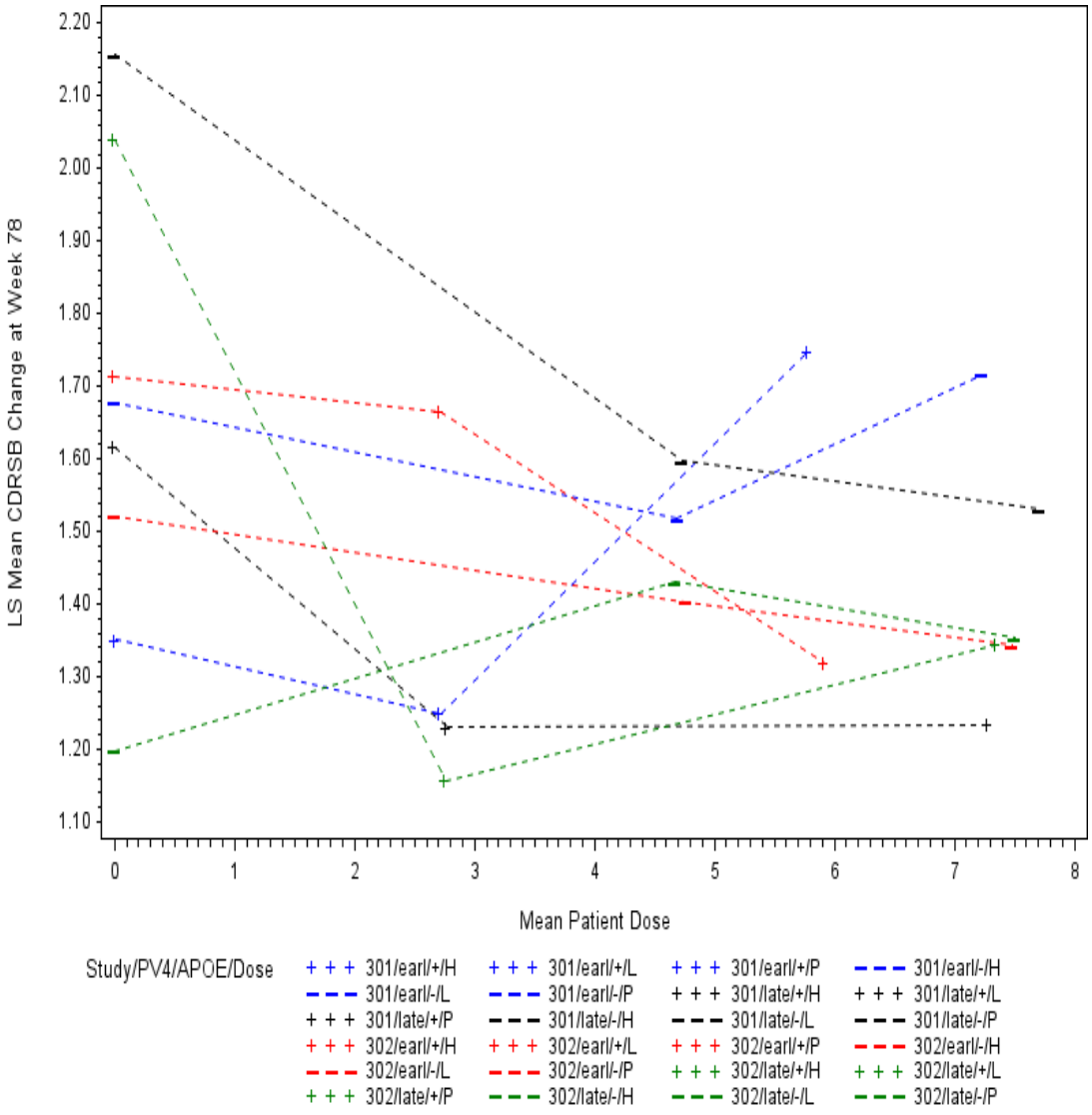


Figure 17 is a similar exploration of dose response at the group level, additionally showing the corresponding placebo responses. The plotting symbol is - for APOE non-carriers and + for carriers and the x-axis identifies the dose since there are two doses for low and two for high specific to carriers and non-carriers except for post-PV4 high which has the highest dose of titration to 10 mg/kg, common to both carriers and non-carriers. The dashed lines joining the symbols are intended to help to discern the dose response. One can observe that there were two outlying placebo response LSmeans, one was for study 302 late or post-PV4 carriers and the other for study 301 late or post-PV4 non-carriers and the dose response between low and high is hit or miss.

Figure 17 Study 302 Exploration of Group Level Dose Response by APOE and Pre-PV4 or Post-PV4 Status



3.2.1.4.2.2 Biomarkers and Limited Correlation between Clinical and Biomarker Changes

The final SAP for biomarker analysis is dated 2020; well after unblinding. The SAP specified before unblinding (12 Sep 2018) stated that correlations based on the pooled-study data would be considered primary. It planned Pearson and Spearman correlations both unadjusted and correlations adjusted for baseline CDRSB and baseline biomarker and also to examine correlations with the biomarker at Week 78, as well as Week 26. The prespecified SAP states that the correlations were to be done by treatment group, which differs from the post-unblinding biomarker SAP which suggested pooling the low and high dose groups.

In the PET biomarker substudy (N=442 subjects from 302) overall 29% of placebo and 22% of the high dose group were missing Week 78 and of those who had the opportunity to complete Week 78 12% of the 99(/140) placebo and 12% of the 113(/145) high dose were missing the Week 78 assessment of composite SUVR. The composite SUVR change from baseline was significant at Week 78 (-0.28 [S.E.= 0.01]) as well as Week 26 (-0.09 [S.E.=0.01]). The low dose was also significant (-0.08 at Week 26 and at Week 78 -0.18 [S.E.=0.01]). Placebo LSMean composite SUVR at Week 78 was 0.0003 [S.E.=0.008] in 301 and +0.0158 [S.E.= 0.01] in 302.

The PET substudy is not directly randomized or balanced. In completers of the PET substudy there were 12% more aged 71-80 in the high dose than placebo. 9-10% more age 71+ in the high dose and a 10% imbalance in APOE. There are 6% more aged 71-80 in the high dose than placebo in the PET population overall (this age group was the group with the highest apparent estimated effect in study 302). Mean CDRSB at baseline in the PET subgroup was 2.39 in study 302 as compared to 2.51 in the high dose overall; in 301 it was also 2.39 in the PET subgroup but 2.40 in the high dose overall.

If these PET SUVR amyloid changes are meaningful why is the biomarker change positive in 301 (N=544 total: High vs. Placebo = -0.07 [S.E.=0.008] at Week 26 and -0.24 [S.E.=0.01] at Week 78) and the biomarker shows dose dependence (low dose -0.07 [S.E.=0.008] at Week 26 and Week 78 -0.17 [S.E.=0.01]) but the clinical change in CDRSB is not significant and the low dose is numerically better than the high dose on CDRSB change from baseline at Week 78 in study 301? Unadjusted Pearson correlations of Week 78 changes from baseline between the biomarker and CDRSB within the high dose group are 0.135 for study 301 and -0.036 for study 302 (see Figure 18 which includes a local regression curve to help the eye assimilate the data points). The unadjusted Pearson correlation with the biomarker for the on-face positive study 302 is in the wrong direction. For example, the high dose patient with the largest decrease in SUVR Amyloid beta Composite with reference in the Cerebellum -0.79 had a Week 78 CDRSB that was a 3.0 point increase (worsening) from baseline. The high dose patient with the second largest Week 78 decrease in Amyloid beta (-0.68) via PET Composite SUVR with reference in the Cerebellum had a week 78 CDRSB, representing an increase of 0.5 from baseline. Pooling

studies 301 and 302, the correlations adjusted for baseline CDRSB and baseline cerebellum SUVR for high dose were 0.145 (p=0.0375) and 0.145 (p=0.0376). However, for study 302 only (N=99), the baseline adjusted Pearson correlation was 0.104 (p=0.3090) and the adjusted Spearman correlation was 0.130 (p=0.2058). For 301 (N=110) the corresponding adjusted correlations (and nominal p-values) were 0.135 (0.1650) for Pearson and 0.074 (0.4455) for Spearman. In summary, for high dose patients the Week 78 CDRSB change from baseline and the Week 78 PET Composite SUVR (with Cerebellum reference region) changes in Aβ from baseline are essentially uncorrelated which raises doubts about disease modification claims and substantial evidence of effectiveness in light of the mixed results of 301 and 302. In fact, while the correlation is nominally significant for pooled high dose group the corresponding regression model with the baseline scores and the biomarker change as predictors of the week 78 CDRSB change has a p-value for the significance of the variance explained by the full model of p=0.188 (not significant). With the high and low dose combined the Pearson correlation was nominally significant in study 302 .168, p=.02 (and Spearman .194), but the Pearson was not significant in study 301, 0.003 p=0.96, and the Spearman correlation of the combined doses was in the wrong direction, i.e., a slightly negative correlation, -0.04, in study 301. Furthermore, in study 302 the low dose had a numerically bigger correlation than the high dose (adjusted Spearman correlation: 0.21 for low dose vs. 0.13 for high dose and 0.16 vs. 0.10 for adjusted Pearson correlation) further complicating the interpretation if one tried to argue that these correlation magnitudes were meaningful. Note that the p-values presented for correlation should not be overinterpreted, they are not very meaningful because they are for exploratory tertiary analyses and are not adjusted for multiplicity.

When adjusted for all explanatory variables in the primary analysis model the within patient correlation for the combined doses between week 78 CDRSB change and week 78 composite SUVR biomarker change is estimated at 0.087 p=0.24, not significant.

If the Week 78 SUVR biomarker change is added to the primary analysis model for CDRSB it is found not to add anything significant to the model p=0.8840 (week 26 biomarker change p=.7875). Therefore, adjusting for the other prespecified model covariates there is essentially no correlation between CDRSB change and change in the biomarker composite SUVR with cerebellum reference.

The sponsor’s mediation analysis also showed Week 78 SUVR biomarker change explained a numerically higher proportion of clinical endpoint treatment effect for low dose than high: 36% low vs 33% high and the 95% bootstrap confidence intervals did not exclude 0% (p 125 of ise-appendix g6). Thus, there is no evidence that the SUVR change is a surrogate for clinical change. Examination of the mediation analysis model suggested that neither baseline biomarker alone is a significant predictor of CDRSB change at Week 78 p=.5809; nor are both baseline and week 78 biomarker when included in the model together (baseline biomarker p=0.3470 ,Week 78 biomarker p=0.3955).

Figure 18. Assessing Correlation of Amyloid Pet and CDRSB in High Dose at Week 78

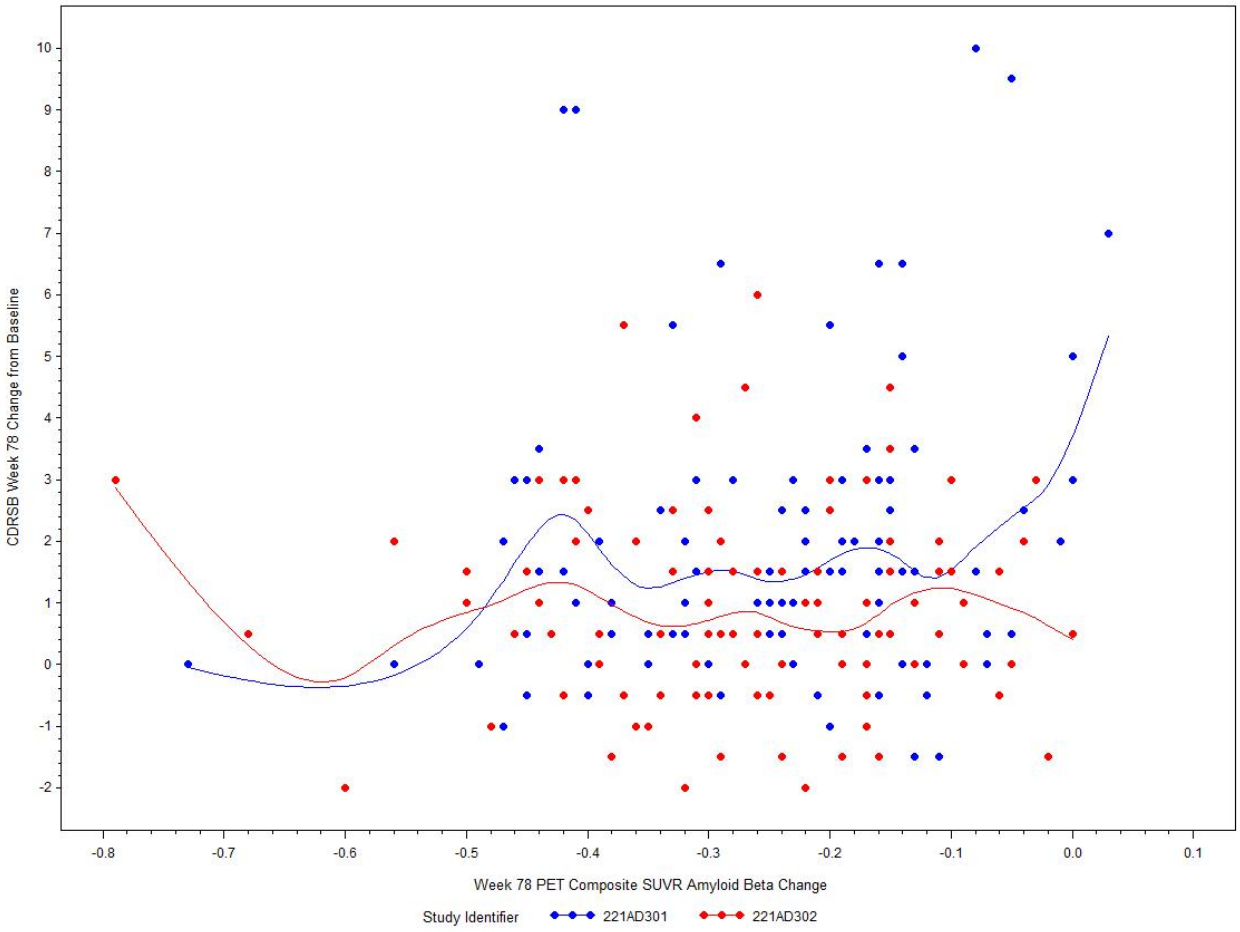


Table 11 presents both baseline-adjusted and baseline-unadjusted Pearson and Spearman correlations between Week 78 CDRSB and Week 78 Composite SUVR by the various possible combinations of Study and Treatment groups. Of the 36 estimated correlations, 21 of them were either less than 0.1 or negative. The largest point estimate of the correlation is for the low dose and this is only 0.22. Thus, while a few of these correlations are nominally significant, the magnitude of the correlation is still rather small and questionable for its meaningfulness.

Table 11 Baseline Adjusted and Unadjusted Pearson and Spearman correlations between Week 78 CDRSB and Week 78 Composite SUVR by Study and Treatment groups

	301 High (N=110)	302 High (N=99)	301 Low (N=136)	302 Low (N=90)	301 Pooled doses	302 Pooled doses	High Pooled study	Low Pooled Study	Pooled studies
Pearson correlation	0.135 (0.160)	-0.036 (0.726)	0.009 (0.921)	0.165 (0.120)	0.026 (0.681)	0.105 (0.150)	0.084 (0.225)	0.083 (0.213)	0.066 (0.167)
Adjusted Pearson	0.135 (0.165)	0.104 (0.309)	-0.027 (0.754)	0.158 (0.142)	0.003 (0.960)	0.168 (0.021)	0.145 (0.038)	0.063 (0.350)	0.079 (0.102)
Spearman correlation	0.107 (0.265)	-0.005 (0.960)	-0.004 (0.966)	0.223 (0.034)	0.003 (0.957)	0.129 (0.078)	0.069 (0.319)	0.085 (0.203)	0.061 (0.201)
Adjusted Spearman	0.074 (0.446)	0.130 (0.206)	-0.049 (0.574)	0.211 (0.048)	0.042 (0.512)	0.194 (0.008)	0.145 (0.038)	0.054 (0.426)	0.075 (0.121)

3.2.1.4.2.2.1 Study Group Level Correlations

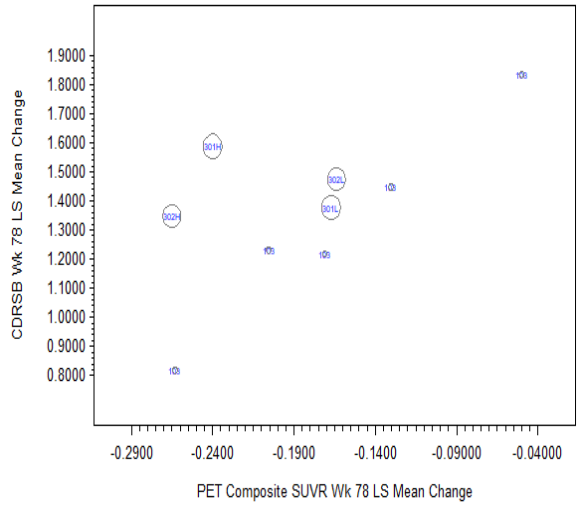
Figure 19 shows study group level correlations between CDRSB changes and Composite SUVR changes both at Week 78 non-placebo subtracted on the left (i.e., actual Aducanumab LSMeans) and placebo subtracted on the right (i.e., differences between Aducanumab and placebo LSMeans). Note that the differences within the same study are correlated due to sharing the same placebo and secondly, the placebo LS means are variable between studies but assumed equal which is obscured by the placebo differencing. We can see that while the 301 high dose appears to be an outlier in the placebo differenced version, it appears less of an outlier and the 10 mg/kg group (high dose 301/302 equivalent) in study 103 appears more of an outlier on the left in the non-placebo differenced version. The right figure should be interpreted cautiously when the placebo means are not very consistent between studies as is the case here, and at least part of the reason why the 301 high dose appears to be an outlier is because among the 3 studies, the placebo LSMean change in CDRSB was least favorable for drug comparison in study 301.

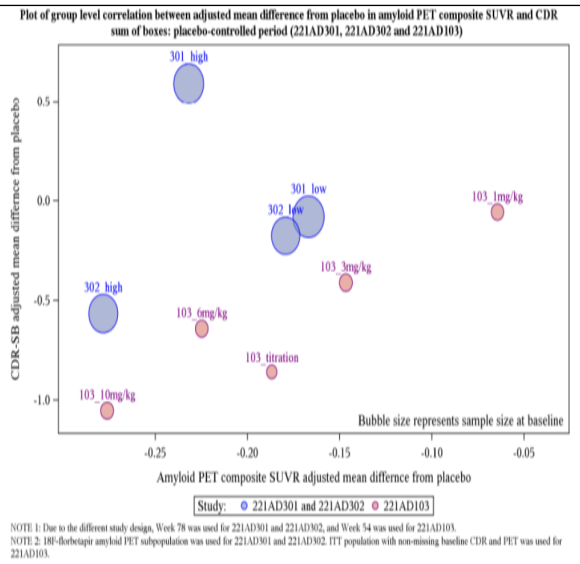
The estimated group level Pearson correlation weighted by study size are .18 for non-placebo subtracted and .19 for placebo subtracted, respectively. Weighting by study size in the correlation analysis is necessary since otherwise the different sample sizes by study result in non-constant variances across the 9 group level means. The unweighted analysis would also give

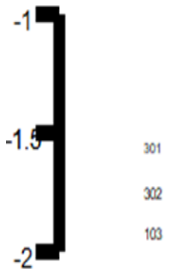
excessive weight to study 103 estimates relative to those of 301 and 302 when study 103 estimates are much less reliable due to their much smaller sample size.

Figure 19 Placebo Differenced and non-Placebo-differenced Aducanumab LS Means Correlations between CDRSB and SUVR at Week 78

Non-placebo subtracted Aducanumab Group LS Means	Placebo subtracted Aducanumab Group LS Means	Placebo LS Means by Study







To summarize, at the group or study level, the correlation between CDRSB change and SUVR change LSMeans at Week 78 (not placebo subtracted) is only r=0.18 when adjusted for study sizes. Furthermore, as mentioned earlier, the proportion of the Week 78 clinical treatment effect in CDRSB explained by Week 78 change in SUVR was only 33% for the high dose and the corresponding 95% confidence interval did not exclude 0% explained. Thus, it is not clear that Week 78 change in SUVR predicts change in Week 78 CDRSB in a meaningful way and there is no compelling evidence that Week 78 change in SUVR is a surrogate. We note that the SUVR biomarker endpoint was an exploratory endpoint with data only collected in a convenience subset of patients (33% with only 18% having Week 78 SUVR data in 302 [in 301, 36% participation and 21% complete SUVR; in 103, 85% participation and 74% complete SUVR])]). Any formal discussion on patient-level or trial-level correlation based on incomplete data should be discouraged and the findings should be viewed as exploratory at best.

3.2.1.4.2.2.2 Other Exploratory Biomarker Correlations at the Patient Level

For the exploratory CSF biomarkers in study 302 only 17/50 (34%) high dose in the CSF substudy completed the Week 78 assessment as compared to (28/55) 51% and 33/64 (52%) for the low dose. With so much incomplete data even within the CSF substudy the CSF results may

not be reliable. Four high dose subjects had Week 78 CSF assessments that were assigned to the long term extension and thus excluded from the Week 78 analysis. The patient level correlations were slightly lower if these patients were included. Secondary biomarkers also did not correlate very well with primary efficacy outcome in the high dose as follows (see Table 12). Aducanumab appears to increase CSFAβ 1-42. Thus, a negative correlation between CDRSB change and CSFAβ change would support an association.

There was a non-significant negative correlation within the high dose between CSF Aβ 1-42 and CDRSB change at Week 78. However, Aβ 1-40 was decreased by Aducanumab compared to placebo but not significantly and the correlation for Aβ 1-40 change and CDRSB change was positive and non-significant. With 301 and 302 pooled (N=30) CSF Aβ 1-42 Week 78 change partial correlations with CDRSB Week 78 change were -0.14 (p=0.47) and -0.25 (p=0.21). With 301 and 302 pooled (N=25) CSFAβ 1-40 Week 78 change correlations with CDRSB Week 78 change were 0.16 (p=0.41) and 0.21 (p=0.29)

Aducanumab appears to decrease CSF ptau and tau protein. Thus, a positive correlation between CDRSB change and CSF tau change would be required for a meaningful association between change on imaging and change on clinical endpoint. The correlation for ptau for the high dose was nominally significant but recall that the completion rate even within the CSF substudy for the high dose was only 34%.

Also, when the doses are combined for study 302 as advocated by the sponsor the ptau correlation is much smaller and not significant (n=45, Pearson=0.248 p=0.1083; Spearman=0.212, p=0.1732). In fact, for the low dose which had a higher completion rate (52% vs. 34%) the Pearson correlation was in the opposite direction of the high (n=28, Pearson=

-0.049 p=0.8136; Spearman=0.050, p=0.8065). For the high dose in study 301 correlation between Week 78 CDRSB change and Week 78 ptau change was much smaller and did not reach nominal significance (n=18, Pearson=0.277 p=0.2999 ; Spearman=0.380 p=0.1463).

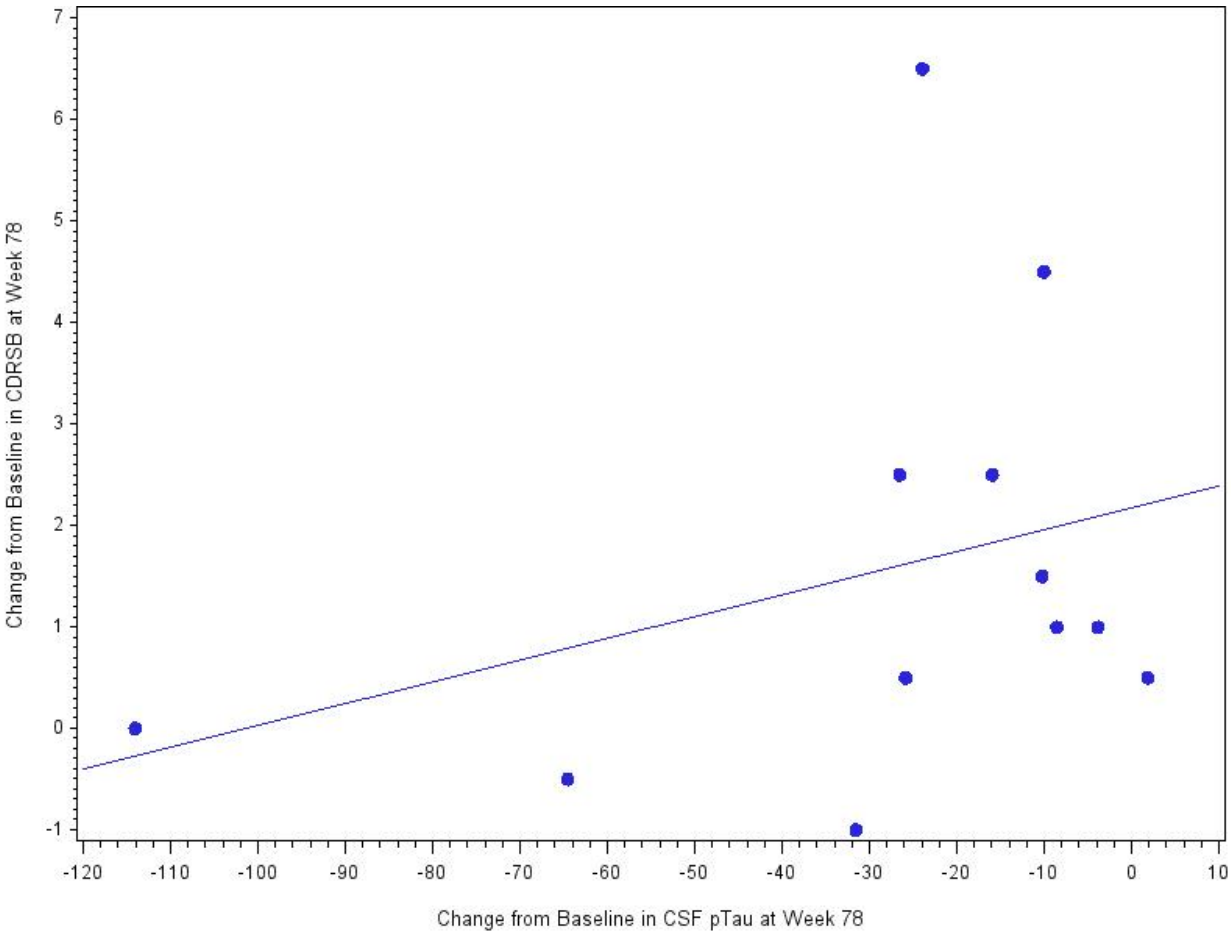
The correlation with tau protein changes and CDRSB changes at Week 78 within the high dose alone in study 302 was not nominally significant.

Table 12 Study 302 High Dose CDRSB Week 78 Change Correlations with Week 78 Biomarker Changes

	Pearson partial (p-value)	Spearman (p-value)	Anticipated direction of correlation to support association
Correlation with CSF Aβ 1-40 n=13	0.05 (0.86)	0.32 (0.26)	none
Correlation with CSF Aβ 1-42 n=13	-0.24 (0.42)	-0.38 ,(0.19)	negative
Correlation with CSF ptau n=12	0.73 (0.002)	0.66 (0.007)	positive
Correlation with CSF tauprot n=11	0.44 (0.099)	0.22 (0.43)	positive

Figure 20 shows Correlation between Week 78 CSF Ptau and Week 78 CDRSB change within the High Dose group of Study 302 (with a fitted regression line). The Week 78 CDRSB change from baseline is identified by the y-axis and the Week 78 change in ptau is identified by the x-axis. The correlation is weak, e.g., the patient with the best CSF change did not improve on the CDRSB and was within 0.5 CDRSB points of the patient with the worst CSF ptau change at Week 78.

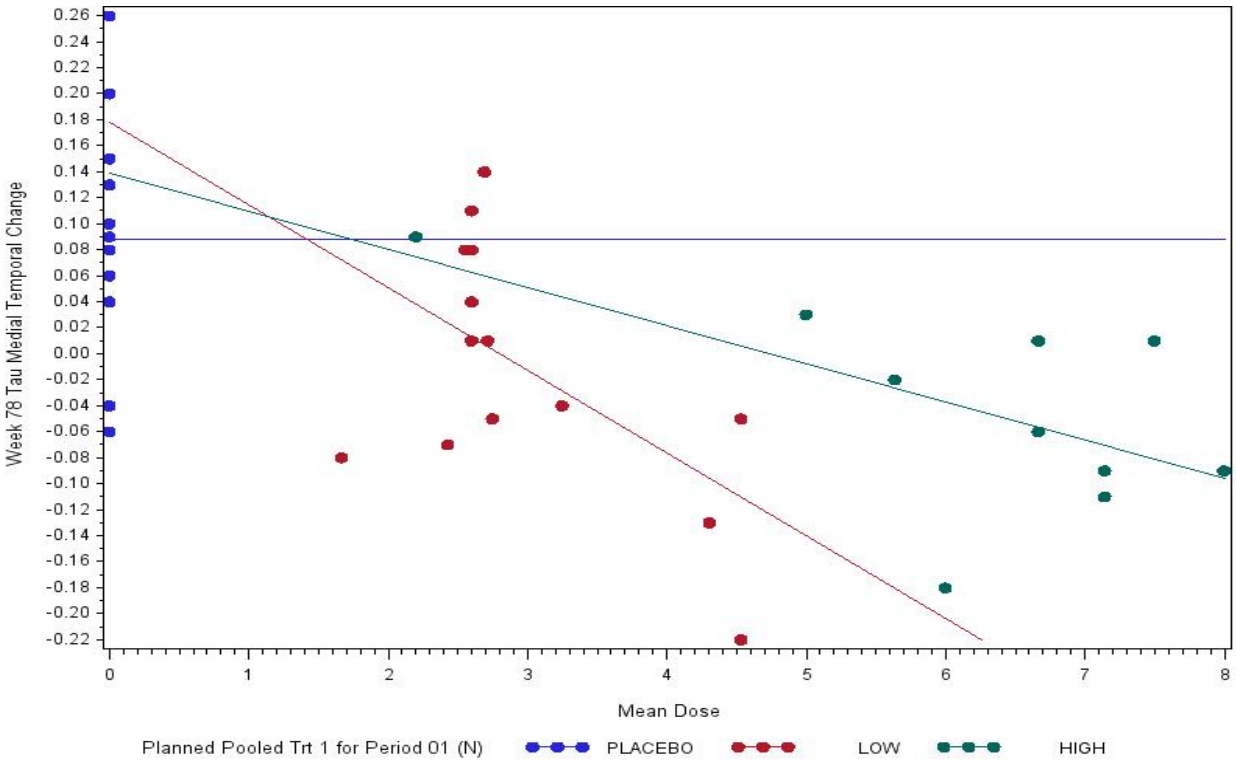
Figure 20. Correlation between Week 78 CSF PTau change and Week 78 CDRSB change within High Dose in 302



The sponsor also highlighted changes in the non-primary biomarker tau medial temporal SUVR at 78 weeks based on just 12 placebo, 14 low dose, and 11 high dose patients. It is important to observe that in this small subgroup the low dose had a stronger slope/correlation between tau medial temporal SUVR and mean dose than did the high dose (Figure 21 which shows a fitted regression line for each dose 1 (red)=Low ; 2 (green)=High). This is only 23 patients total (excluding the low dose) and the Tau biomarker is downstream in the disease process from the drug's target of Amyloid. Therefore, without a compelling correlation between amyloid change

and CDRSB change this possible relationship further downstream seems dubious. In fact, these two groups of 12 and 11 patients are not very well balanced. The baseline CDRSB was 2.75 for the 12 placebo and 2.27 for the 11 high, 8 vs. 18 % had Mild disease, 33% vs. 55% used concomitant medications at baseline, average age was 69 vs. 76 and 8% vs. 0% were from study 302. In short, this is a very small convenience sample and it's not clear that the groups were even comparable at baseline. There was no effect overall in study 301, the parent study from which most of these 23 patients came, so it is additionally hard to believe this tau result in a tiny substudy. Furthermore, the correlation between tau and CDRSB change at week 78 in the high dose (all measured after the futility announcement) is not significant-=-0.35 p=0.5272(n=6 with non -missing Week 78 CDRSB). This correlation should have been positive (>0) in order to support the high dose causality, i.e., reduction of tau translating to long term clinical change in CDRSB.

Figure 21. Week 78 Change in Medial Temporal Tau by Mean Dose (Study 301 and 302)



3.2.2 Study 301

3.2.2.1.1 Sponsor's Results

This was covered above. Please see section 3.2.1.4.1.

3.2.2.1.2 Reviewer's Results

Study 301 was negative overall with the high dose actually being numerically worse than placebo (estimated difference =+.03 [S.E.=.150], 95% C.I.= [-.262,.328], p=0.8). It can rule out a high dose drug effect greater in magnitude than .262 with 95% confidence. The sponsor tried to find similar patients in 301 to 302 and highlighted an apparent high dose trend in those who were under the protocol version 4 amendment. However, there is a multiplicity problem with this (overall the trial was negative, this looks at a subset that was not even prespecified, and there is only a partial randomization: it is only part of the process and this could create important exclusions, that again were not prespecified). Since the sponsor made a lot of effort to find 301 subjects like 302 subjects this reviewer evaluated the consistency post-PV4 in study 301.

Some patients randomized as late as Sept 2017 (by which one would have to be randomized in order to have the opportunity to complete by March 20, 2019) were not under protocol version 4 for some reason (e.g., patient consent/ IRB/site acceptance, etc.) , thus did not get 10 mg/kg dose or at least not for the maximum possible time if they were APOE+. The non-PV4 patients seem to have different baseline characteristics than those randomized under protocol version 4 in the same time frame. Once again, this means the effect of raising the dose in APOE+ is confounded with enrollment changes over time in baseline characteristics (e.g., country and baseline disease stage as shown in Figure 1 for study 302 and the result is similar for study 301: 26% non-OTC were Mild vs. 16% OTC and US region decreased from 51% in OTC to 36% in non-OTC). The pre-PV4 vs. post-PV4 grouping is slightly different from the OTC vs . non-OTC grouping but both show some differences, e.g., 61% post-PV4 vs. 53% pre-PV4 used concomitant AD medications at baseline and 13% vs. 6% were Asian.

In the APOE+ subgroup, which seems to drive study 302, the high dose was in the wrong direction overall in study 301: CDRSB Week 78 LS Mean High 1.51 vs. Placebo 1.44, thus APOE+ high vs. placebo difference is +0.070 (S.E.= 0.180), p= 0.6971. [in APOE - the difference is -0.058 (S.E.=0.272)].

In the subset of data prior to protocol amendment 4 (N=815), overall, the high dose group was nominally significantly worse at Week 78 on CDRSB than the low dose (p=0.0303) [and worse than placebo at the .15 significance level].

In the subset of data after protocol amendment 4 (those consenting to it by day 113 of their follow-up), the estimated high dose treatment difference at Week 78 was in the right direction: N=787 subjects:

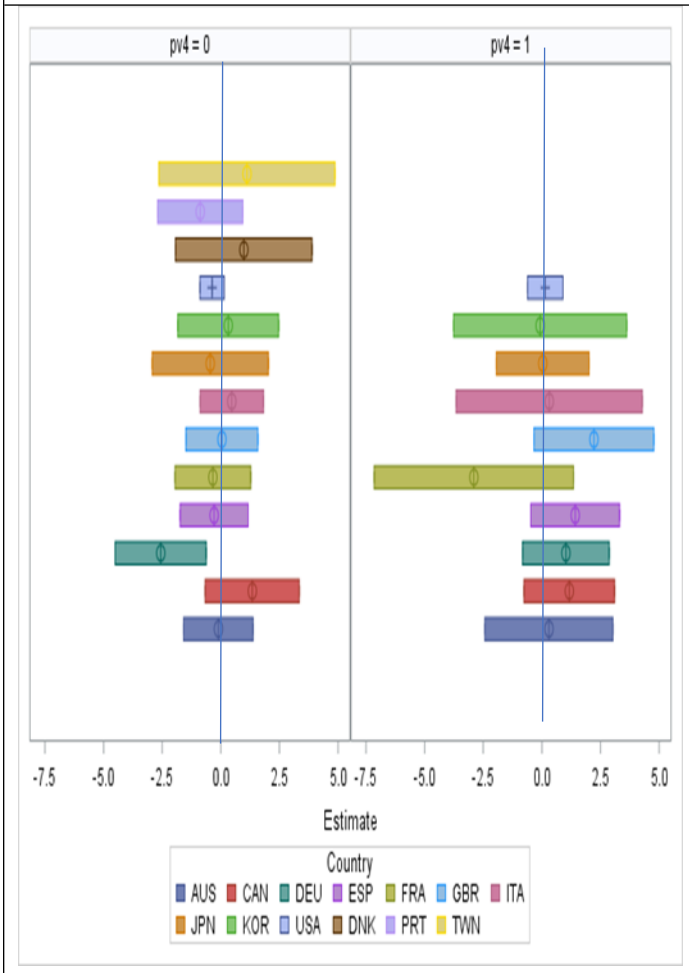
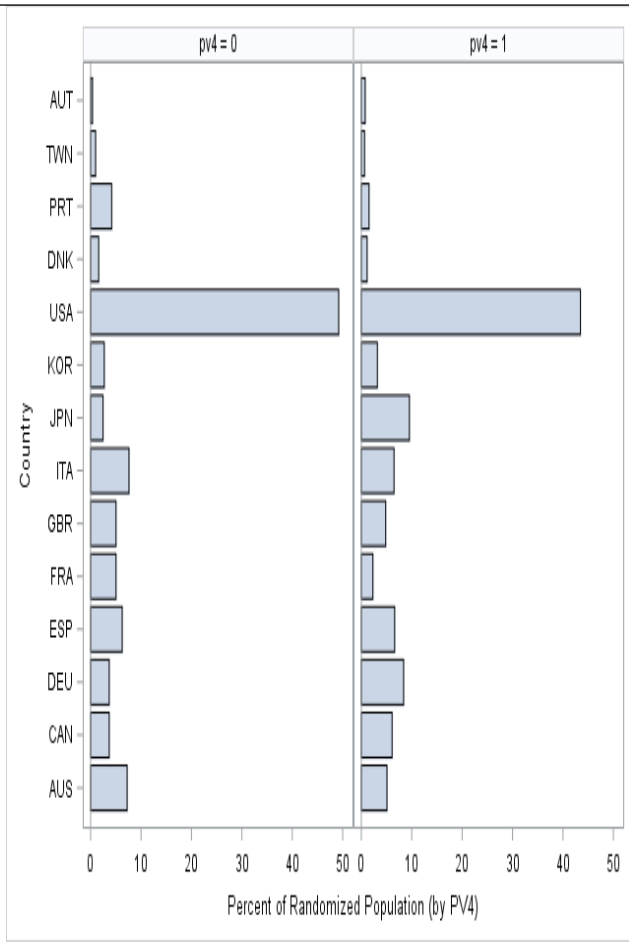
-0.487 [S.E.=0.270], p=0.0724.

However, in this post-PV4 subset there was a significant interaction between Treatment group, Country, and Visit: country*visit* treatment (country main effect p=.056; 3 way Country by Visit by Treatment interaction: p=0.07; all country terms p=0.0351; all interaction effects beyond region main effect : p=0.0625 [101 numerator degrees of freedom]). If the more coarse, prespecified Region effect is used instead of Country in the model the 3 way interaction Region*Visit*Treatment p-value is 0.0139. Thus, whether adjusting for prespecified Region or Country, the treatment differences appear to vary significantly by Region or Country in the post-PV4 subset. The country interaction is even more compelling in the APOE+ subgroup (main p=0.0097; 3 way interaction p=0.0548).

Figure 22 shows the high vs. placebo estimated differences (+/- 1 Std Error) by Country in pre-PV4 and post-PV4 subsets with positive differences favoring the high dose in the figure.

Even in the post-PV4 subset, the 301 high dose effect relative to placebo in the US, the largest enroller, is very small (and would not be nominally significant). Also, the post-PV4 high dose CDRSB effect in Japan is virtually zero although it was big in study 302.

Figure 22. Study 301 High Treatment Effect Estimates at Week 78 pre-PV4 and post-PV4

Hi Dose vs Placebo Week 78 CDRSB effects by Country Pre-PV4 and Post-PV4	Enrollment by Country Pre-PV4 and Post-PV4
	

The estimated high dose effect in the post-PV4 subset also seems to vary significantly by age group (age group main effect test p=0.0003; agegrp*visit*trt interaction test p=0.0352). Only the oldest group, >75 accounting for 34% of post-PV4 patients, showed a nominal effect at Week 78 on CDRSB for the high dose and the <65 subgroup (accounting for 21%) was in the wrong direction (Table 13).

Table 13. Study 301 Post-PV4 subset High Dose vs Placebo CDRSB Treatment Effect Estimates by Age Group

Age Group	LSMean Difference High Dose -Placebo at Week 78 +/- Std. Error
<65	+.16 +/- .6
65-74	-.24 +/- .4
≥75	-1.2 +/-.5

In study 301, the high dose effect in APOE- (non-carriers) is numerically better post-PV4 than in APOE+ (Table 14), but the dose increase for the high dose was in APOE+ only and pre-PV4 in APOE- there was essentially no effect (compare

Figure 23 blue bars at Week 78).

Table 14. Study 301 Post-PV4 High -Placebo Estimated CDRSB Difference at Week 78 by APOE subgroups

APOE subgroup	Comparison	Estimated Difference (LSMean)	Std. Error	Nominal p-value
APOE+	High - Placebo Week 78	-0.43	0.33	0.186
APOE-	High - Placebo Week 78	-0.62	0.49	0.202

Figure 23 shows CDRSB profiles in Study 301 APOE - (non-carrier) stratum in the post-PV4 subset. The most striking change from pre-PV4 to post-PV4 is the fact that Placebo worsened markedly at Week 78 as compared to pre-PV4. Also, comparing to the low dose, post-PV4 (red) high and low groups are almost the same at Week 78.

Figure 23. Study 301 Placebo, Low Dose, and High Dose CDRSB Profiles in Study 301 APOE non-carriers Stratum

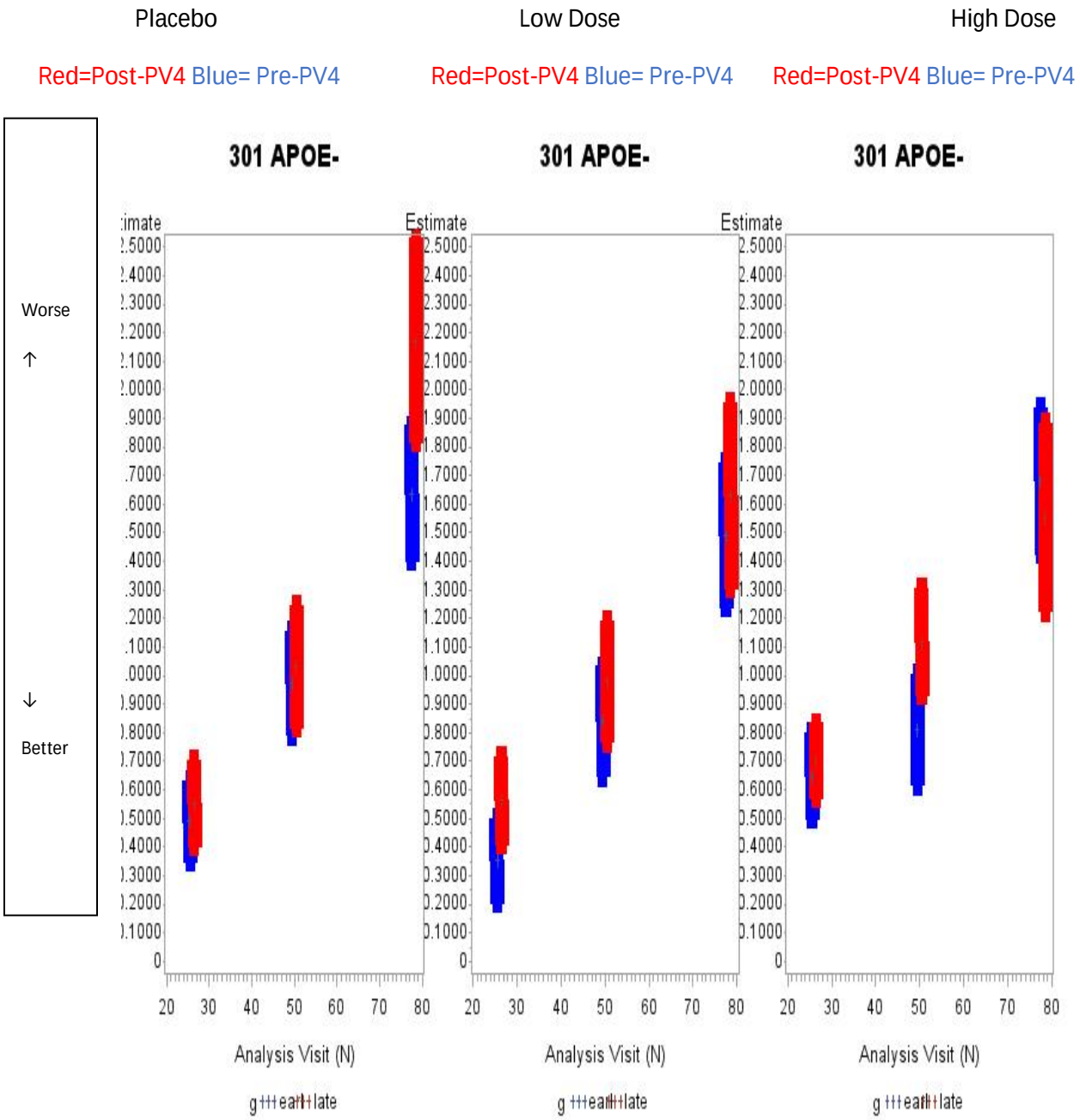
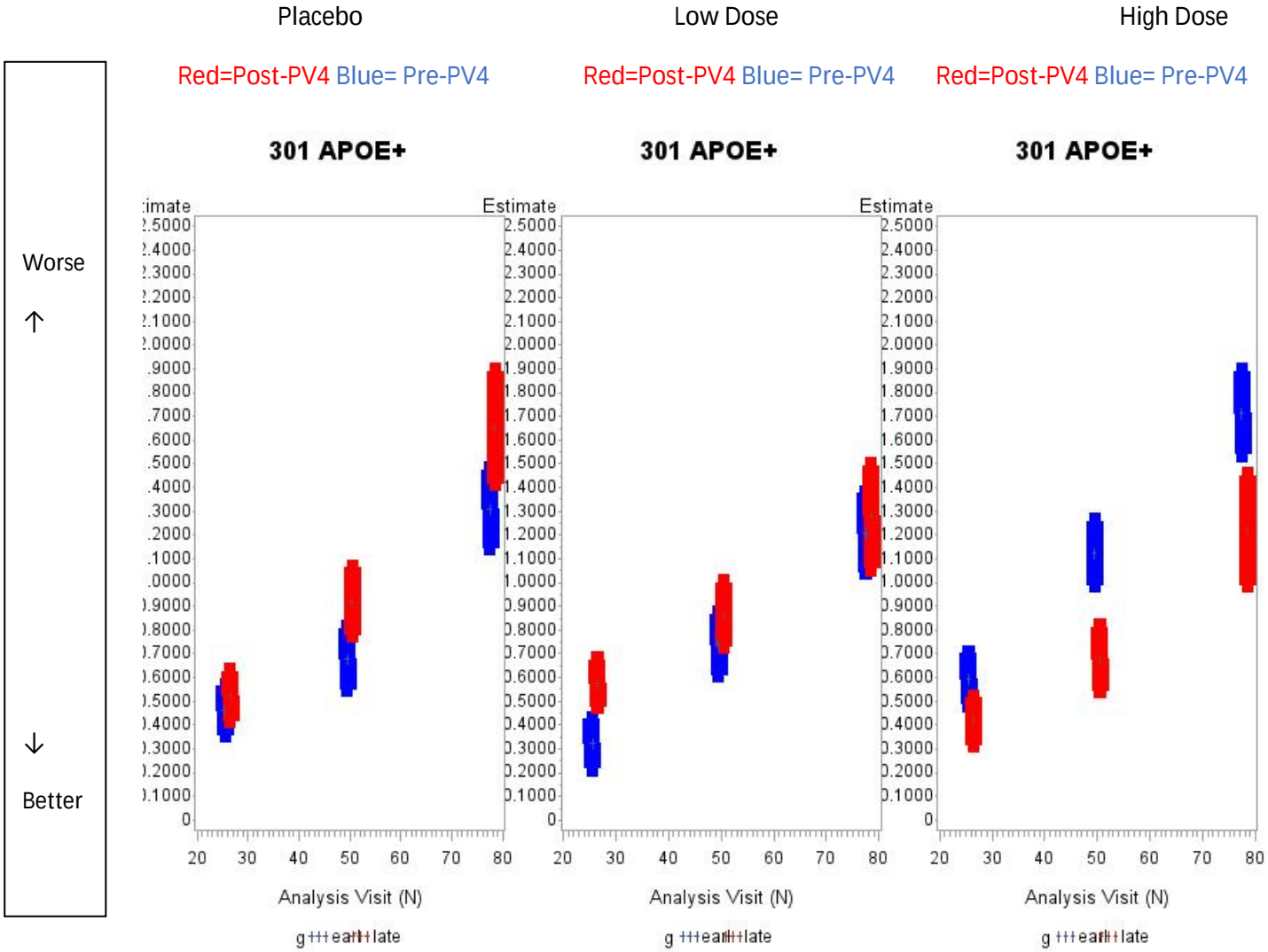


Figure 24 shows CDRSB profiles in Study 301 APOE + (carrier) stratum. While the high dose, on the right, improved post-PV4 relative to pre-PV4, placebo, on the left, also worsened from pre-PV4 to post-PV4. Furthermore, comparing to the low dose, post-PV4 (red) high and low are almost the same, despite the high dose increase.

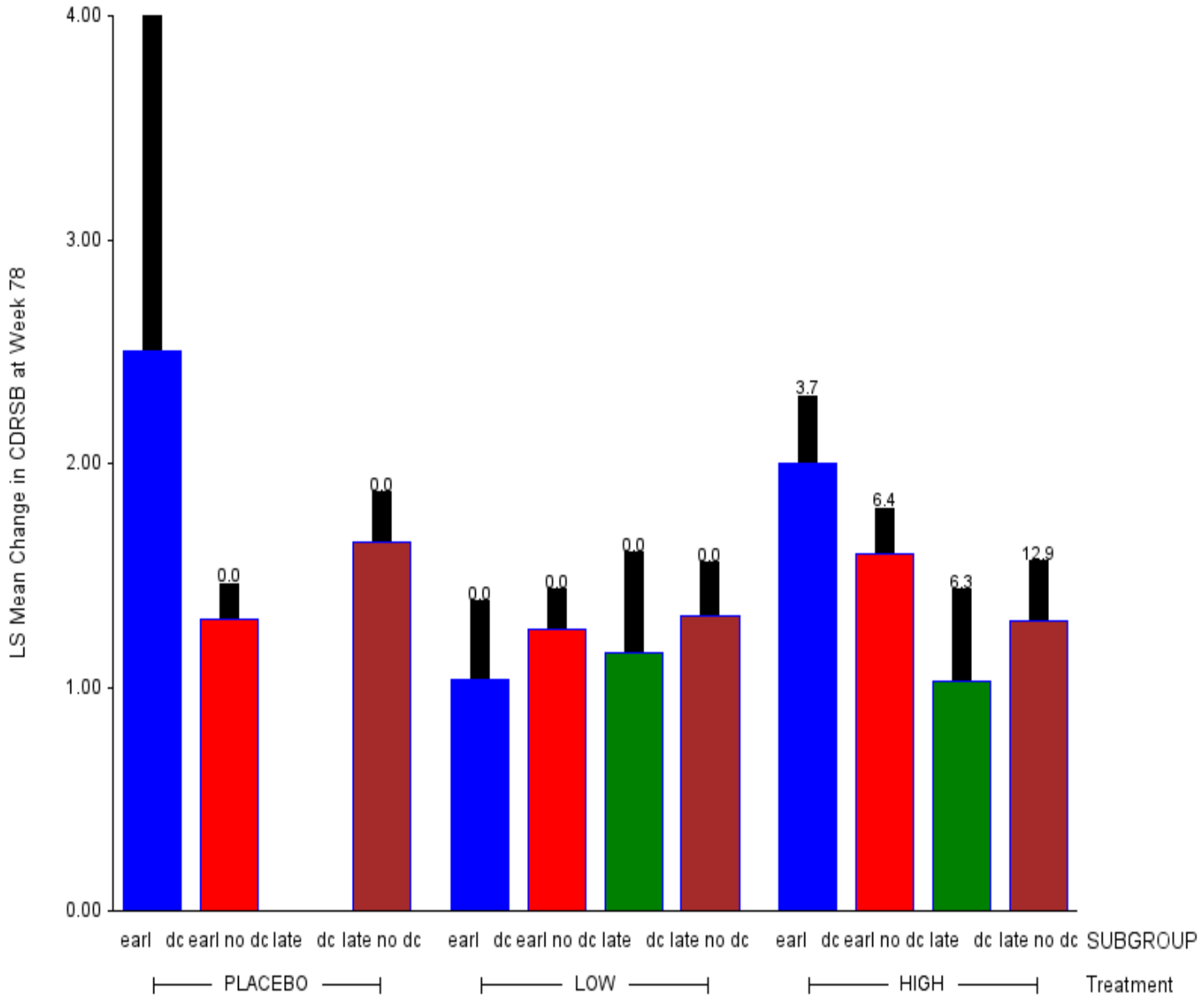
Figure 24 Study 301 Placebo, Low Dose, High Dose CDRSB profiles in Study 301 APOE carrier Stratum



The following graph shows the LS Mean Changes in CDRSB at Week 78 for the APOE+ stratum by Early vs. Late and Dose Reduction due to ARIA adverse event status. The average number of 10 mg/kg doses completed by Week 78 is the value shown above the bars. It appears that the Late non-dose-reduced High Dose subgroup (brown) had significantly more 10 mg/kg doses, as expected, but the LS Mean is numerically worse than for the Late dose-reduced High Dose subgroup (green). The sample sizes are admittedly small but it is not clear why in Figure 25,

as in Figure 10 for study 302 (June 2019 data showed the same pattern as here for 301), the post-PV4 (late) subgroup not requiring dose modification or titration slowing, with more 10 mg/kg doses is numerically worse than the corresponding high dose subgroup that required dose modifications or slowing due to ARIA adverse events. There was some real and more potential unblinding in the latter subgroup, although the rating physician was not the same as the treating physician. There also appears to be very little difference between the low (middle brown) and high dose (right brown) LS means post-PV4 after the high dose had been allowed to increase, again calling into question the importance of the late dose increase for high dose APOE carriers.

Figure 25. Study 301 LS Mean Change from Baseline in CDRSB at Week 78 in APOE+ by PV4 and ARIA dose modification



Note: earl=Pre-PV4 late=Post-PV4; dc= dose titration slowing or reduction due to ARIA

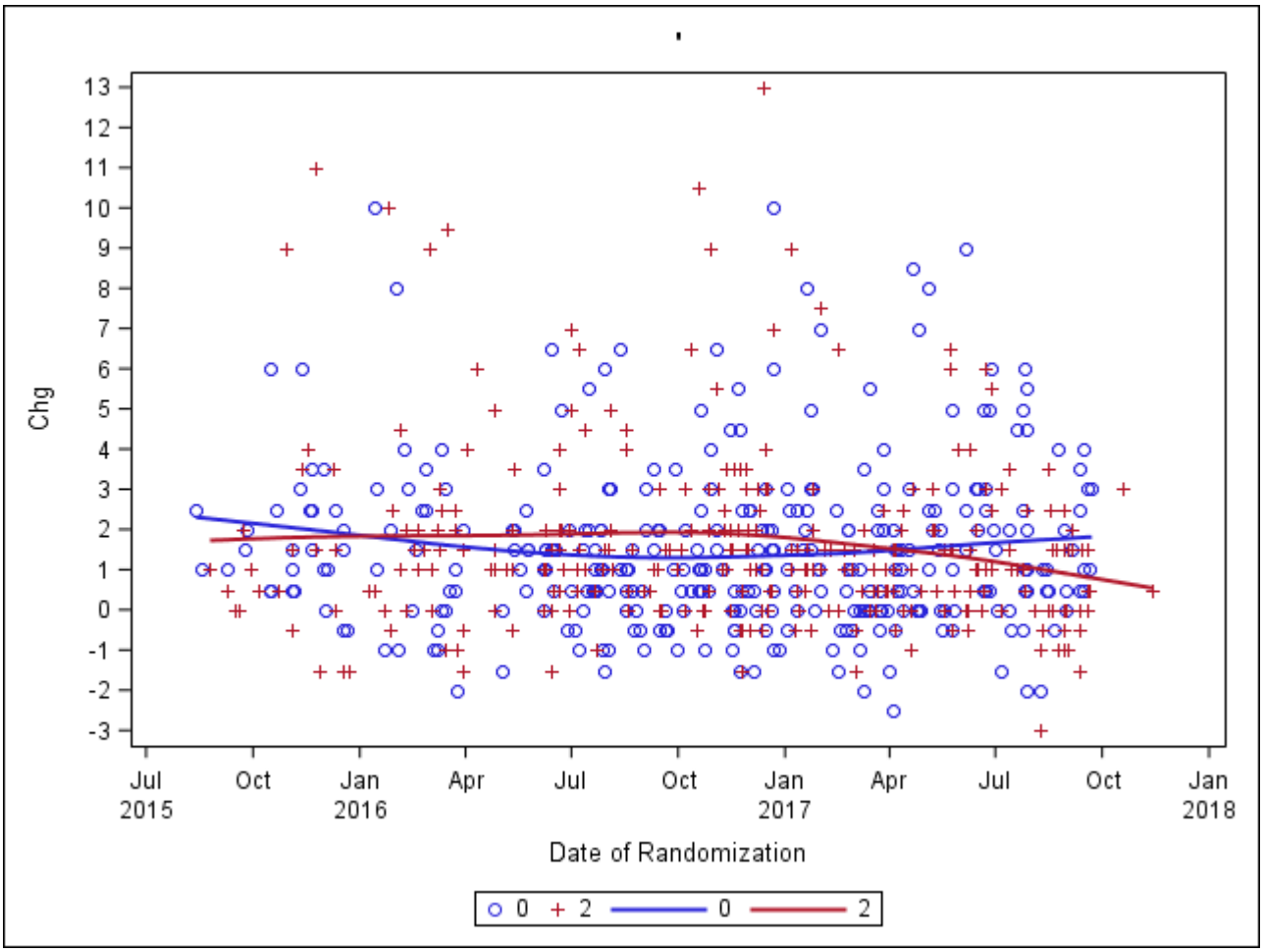
Note that subjects with ARIA-E or ARIA-H adverse events had additional MRIs every 4 or 2 weeks, respectively, including a biomarker and PK sample for ARIA-H, until resolution, which being beyond the normal schedule could have resulted in unblinding and the sponsor Protocol version 1 stated that "Subjects should complete all scheduled clinic visits for assessments and, in addition, have an unscheduled visit for an MRI and MOCA approximately every 4 weeks until the ARIA-E has resolved per the centrally read MRI. Investigators, study site staff (except for the designated unblinded pharmacist/technician), and study subjects will be blinded to the subjects' randomized treatment assignment for the placebo-controlled period."

The central MRI reading center was to report incident cases of ARIA-E and ARIA-H to both the Sponsor and the Principal Investigator within a specified time after observing the finding on MRI per the imaging manual procedures. All cases of ARIA were to be reviewed by the Sponsor and the Principal Investigator; decisions on dosing continuation, interruption, or discontinuation were to be based on clinical symptoms, and the MRI information provided by the central reader. The study documentation further indicated that "Selective roles within study team at Biogen may have access to individual cases of ARIA in order to carry out the operational aspect of the study. They will not perform aggregate summaries or communicate ARIA related data to other team members who do not have access to ARIA data."

Protocol Version 4 was not implemented at the same time across sites and patients had to provide new consent to the new version, so there is not a unique implementation time across the studies and which patients it applies to might have multiple possible definitions. The sponsors definition might not be the only possible definition, e.g., there are APOE + patients who had some 10 mg/kg doses but are classified as non-PV4 by the sponsor because they didn't have the opportunity for all fourteen 10 mg/kg doses. There is a multiplicity issue if one evaluates the PV4 subgroup in addition to the overall study population. PV4 also added a MRI monitoring assessment for ARIA at Week 66. Since this drug related adverse event required some unblinding it could have created the potential for operational bias (even some sponsor personnel were involved in individual dose management decisions). Although they are subgroup estimates and some sample sizes are small one would not expect the high dose managed (decreased or slowed titration) subgroup due to ARIA to have a numerically better Week 78 outcome than the group that did not require dose management due to ARIA yet that is what Figure 25 above shows. It seems to either suggest the possibility of operational bias due to ARIA related unblinding or at least conflicts with the claim that more 10 mg/kg doses lead to more effect vs. placebo.

The study 301 local regression trend in CDRSB change at Week 78 over randomization time for placebo (0 blue) and high dose (=2 red) is shown in Figure 26 as an aid to exploring possible trends in treatment group mean CDRSB changes at Week 78 over calendar time.

Figure 26 **Study 301 local trend in CDRSB change at Week 78 over randomization time**



3.2.3 Study 103

Protocol 221AD103 Version 11 (2018)

The first patient's first dose was administered on 5 October 2012 and the last patient completed week 54 on 11 May 2016. Arms 8 and 9 subjects (titration regimen) were added to the study in protocol version 6 (08Jul2014). The last 10 mg/kg arm subject completed Week 54 on 18 July 2014. The original SAP was finalized on 07 February 2014.

Primary Objective

The primary objective of the study is to evaluate the safety and tolerability of multiple doses of aducanumab in subjects with prodromal or mild AD.

The primary endpoint is:
• Safety and tolerability as measured by incidence of AEs/ SAEs; clinical laboratory test data; vital signs; neurological and physical examination findings; 12-lead electrocardiogram (ECG) data; and brain MRI findings including the incidence of ARIA-E or ARIA-H.
The secondary objectives of this study are:
• To assess the effect on cerebral amyloid plaque content as measured by 18F-AV-45 PET imaging (at Week 26).
• To assess the multiple dose serum concentrations of BIIB037.
• To evaluate the immunogenicity of BIIB037 after multiple dose administration in this population.
The secondary endpoints are:
• Change from Baseline to Week 26 in 18F-AV-45 PET signal in certain brain areas.
• Serum PK of BIIB037 determined by nonlinear mixed effects approach.
• Incidence of anti-BIIB037 antibodies in serum compared to Baseline.
The exploratory objectives of the study include:
• To assess the effect of BIIB037 on the clinical progression of AD.

This is a Phase 1b, multicenter, randomized, double-blinded, placebo-controlled, multiple dose study of aducanumab in subjects with prodromal or mild AD (Mini Mental State Examination [MMSE] scores from 20 to 30). The study was conducted with a staggered, parallel group design, with the first 3 treatment arms (Arms 1-3) conducted in parallel, followed by Arms 4-5 beginning in parallel, followed by Arms 6-7 beginning in parallel, and finally followed by Arms 8-9 beginning in parallel. Arms 1-3 comprised 2 aducanumab arms (1 and 3 mg/kg) and 1 placebo arm; Arms 4-5 comprised 1 aducanumab arm (10 mg/kg) and 1 placebo arm; Arms 6-7 comprised 1 aducanumab arm (6 mg/kg) and 1 placebo arm. Arms 8-9 comprised 1 aducanumab titration arm (1 mg/kg for the first 2 doses, 3 mg/kg for the next 4 doses, 6 mg/kg for the next 5 doses, and 10 mg/kg thereafter) and 1 placebo arm. Titration up to 10 mg/kg in Arms 8-9 was to only be implemented after the DMC's review of all the accumulated safety prior to the first subject receiving 10 mg/kg. Based on subsequent reviews, the DMC was to determine whether titration up to 10 mg/kg in Arms 8-9 should continue. Approximately 188 subjects were to be enrolled in total across approximately 35 centers. See Figure 27 for the study design. Subjects were to receive 14 doses of aducanumab or placebo that was to be administered by IV infusion once every 4 weeks.
Arms 1-3: approximately 80 subjects in total:

- Arm 1: 1 mg/kg aducanumab per dose (30 subjects)
- Arm 2: 3 mg/kg aducanumab per dose (30 subjects)
- Arm 3: Placebo (20 subjects)

Arms 4-5: approximately 40 subjects in total:

- Arm 4: 10 mg/kg aducanumab per dose (30 subjects),
- Arm 5: Placebo (10 subjects)

Arms 6-7: approximately 40 subjects in total:

- Arm 6: 6 mg/kg aducanumab per dose (30 subjects)
- Arm 7: Placebo (10 subjects)

Arms 8-9: approximately 28 subjects in total:

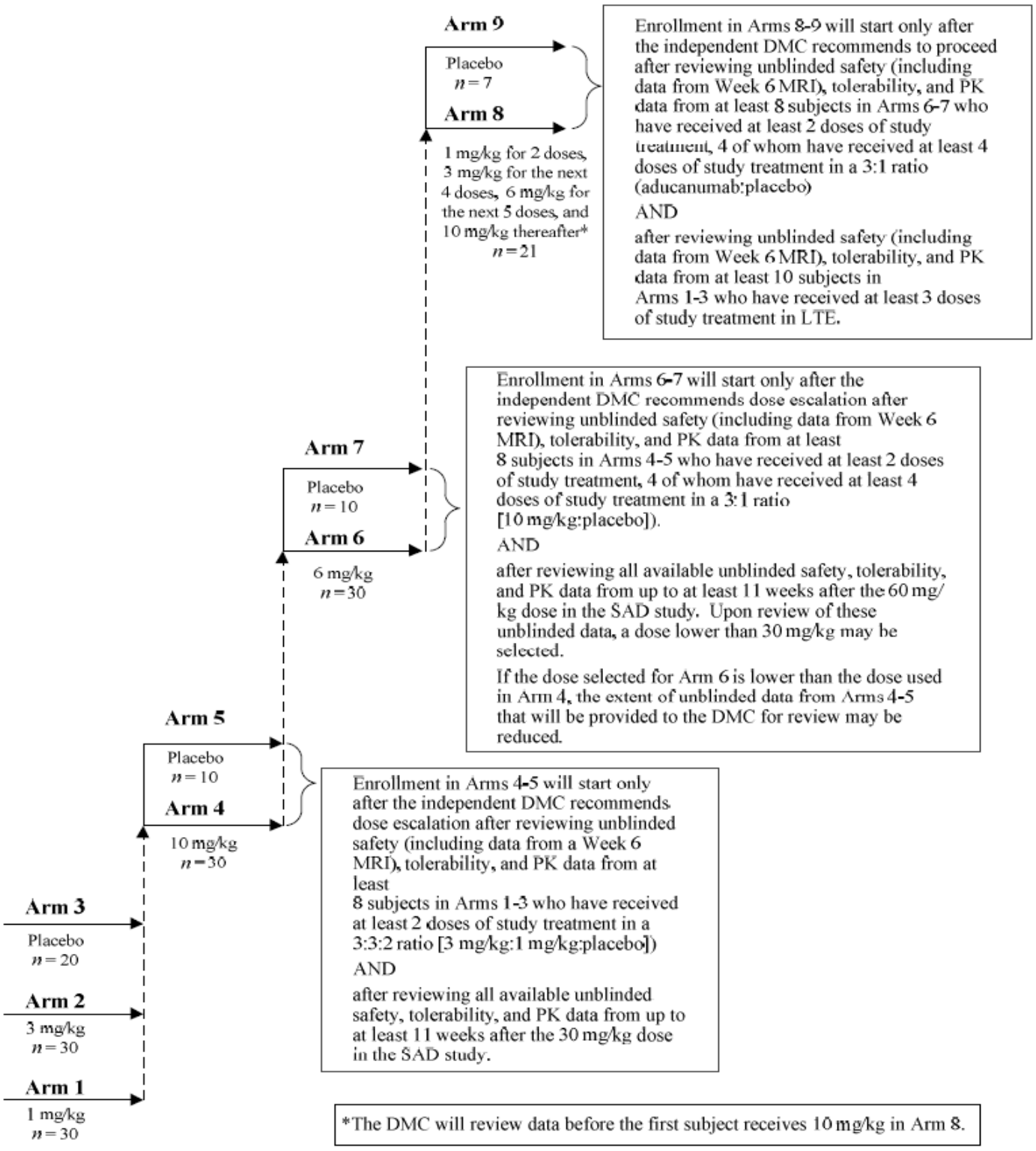
- Arm 8: aducanumab titration arm, 1 mg/kg for the first 2 doses, 3 mg/kg for the next 4 doses, 6 mg/kg for the next 5 doses, and 10 mg/kg thereafter (21 subjects)
- Arm 9: Placebo (7 subjects)

Subjects were to be randomized into each treatment group within Arms 1-3, Arms 4-5, Arms 6-7, and Arms 8-9. The randomization was to be stratified by the ApoE4 status (carrier or non-carrier), with the exception of Arms 8-9, which was to contain ApoE4 carriers only. Enrollment in Arms 1, 2, and 3 was to occur in parallel. Once enrollment in Arms 4 and 5 is open, the enrollment of Arms 1, 2, and 3 and Arms 4 and 5 was to occur in parallel. Enrollment in Arms 6 and 7 was to start once enrollment in Arms 1-5 was completed; enrollment in Arms 6 and 7 was occur in parallel.

Enrollment in Arms 8-9 was to start once enrollment in Arms 6-7 was completed; enrollment in Arms 8-9 was to occur in parallel. The assignment of subjects to Arms 1-3, Arms 4-5, and Arms 6-7 was to be made in a way to ensure the distribution of subjects is comparable between Arms 1-3, Arms 4-5, and Arms 6-7 with respect to the ApoE4 status; the ratio of ApoE4 carriers to non-carriers was to be no more than 2:1 and no less than 1:2. All subjects in Arms 8-9 were to be ApoE4 carriers.

Subjects who met the Long Term Extension (LTE) inclusion/exclusion criteria were to be eligible to enter the LTE for an additional 42 intravenous doses of aducanumab during the first 3 years of the LTE, once every 4 weeks, with the first dose administered approximately 4 weeks after the final dose in the placebo-controlled portion of the study.

Figure 27 Study 103 Staggered Arm Design



Note: Figure copied from page 40 of sponsor's study report

Randomization was to take place across all study sites using a centralized interactive voice/web response System (IXRS).

In Arms 1-3, approximately 80 subjects were to be randomized 3:3:2 of aducanumab 1 mg/kg, 3 mg/kg, or placebo. In Arms 4-5, approximately 40 subjects were to be randomized to a dose of 10 mg/kg aducanumab or placebo in a 3:1 ratio. In Arms 6-7, approximately 40 subjects were to be randomized to a dose of 6 mg/kg aducanumab or placebo in a 3:1 ratio. In Arms 8-9, approximately 28 subjects were to be randomized 3:1 of aducanumab (1 mg/kg for the first 2 doses, 3 mg/kg for the next 4 doses, 6 mg/kg for the next 5 doses, and 10 mg/kg thereafter) or placebo. The randomization was to be stratified by the ApoE4 status (carrier or non-carrier), with the exception of Arms 8-9, which was to contain ApoE4 carriers only.

The protocol allowed for subjects who discontinued study treatment or withdraw prematurely from the placebo-controlled portion of the study to be replaced. Replacement subjects were to be assigned to the same group (i.e., Arms 1-3, Arms 4-5, Arms 6-7, or Arms 8-9) as the subject(s) that withdrew and were to be randomized into a treatment group within Arms 1-3, Arms 4-5, Arms 6-7, or Arms 8-9, as applicable.

Blinding Procedures

This study consists of a randomized, double-blinded, placebo-controlled portion, followed by a dose-blinded LTE with all subjects receiving aducanumab.

For the double-blinded placebo-controlled portion all study staff (except for a designated Pharmacist/Technician and the independent DMC) and subjects were to be blinded to the subject treatment assignments. To maintain the study blind, it is imperative that subject treatment assignments are not shared with the subjects, their families, caregivers, legal representatives, or any member of the site study staff. The Biogen aducanumab study team was to be blinded until the interim analysis on the Week 26 data; information from Arms 1-5 was to be unblinded after all subjects in Arms 1-5 had completed tests and evaluations at the Week 26 Visit, information from Arms 6-7 was to be unblinded after all subjects in Arms 6-7 had completed tests and evaluations at the Week 26 Visit, and information from Arms 8-9 was to be unblinded after all subjects in Arms 8-9 had completed tests and evaluations at the Week 26 Visit. Week 26 information was to only be unblinded to a limited team within Biogen.

Analysis Population

The PD analysis population is defined as all subjects who were randomized, received at least one dose of study treatment, and have at least one post-baseline assessment of the parameter being analyzed.

Methods of Analysis

The change from Baseline to Week 26 and Week 54 in amyloid plaque burden as measured by PET was to be summarized overall and by ApoE4 status for each treatment group and was to be analyzed by analysis of covariance adjusting for baseline amyloid plaque burden and the ApoE4 status to assess the dose response and pairwise comparison with placebo. The analysis may be performed by the ApoE4 status. Due to the exploratory nature of this study, there were to be no multiple comparison adjustments.

Change from baseline in CDR Neurospsychological Test Battery, MMSE, Free and Cued Selective Reminding Test (FCSRT), and Neuropsychiatric Inventory-Questionnaire (NPI-Q)) were to be summarized by treatment group. CDRSB was to be analyzed by ANCOVA adjusting its baseline and ApoE status (as reported by (b) (4)), similarly to SUVR, at Week 26 and Week 54 separately. A sensitivity analysis was to be conducted using the per protocol analysis population. MMRM was also to be performed as a sensitivity analysis. Additional sensitivity analysis was to include treating visit as a continuous variable in the model.

Disease-related biomarkers, change from baseline in morphometric MRI measures, cerebral blood flow (by ASL-MRI), functional connectivity (by tf-fMRI), and glucose metabolism (by FDG PET) were to be summarized by treatment group. Analysis of covariance or its non-parametric equivalent was to be used to analyze these exploratory endpoints.

Interim Analyses

Interim analyses were to be performed for the purpose of planning future studies of aducanumab, and no changes were to be made for this study based on the interim analysis results. A limited team from Biogen was to be unblinded at the interim analysis. Up to 6 interim analyses for placebo-controlled period data may be performed.

- After all subjects in Arms 1-7 have completed tests and evaluations at the Week 26 Visit. Analysis of the primary PD endpoint (change from baseline to Week 26 in ^{18}F-AV-45 PET signal in certain brain areas) will be included in this interim analysis.
- After all subjects in Arms 1-5 have completed the Week 26 Visit
- After all subjects in Arms 8-9 have completed the Week 26 Visit.
- After all subjects in Arms 1-5 have completed the Week 54 Visit.
- After all subjects in Arms 6-7 have completed the Week 54 Visit.

- After all subjects in Arms 8-9 have completed the Week 54 Visit. The Sponsor may perform additional interim analyses.

Reviewer's Comment: The final study report states that to support clinical development plan needs and scientific communications, a series of 7 interim analyses were conducted at pre-established milestones.

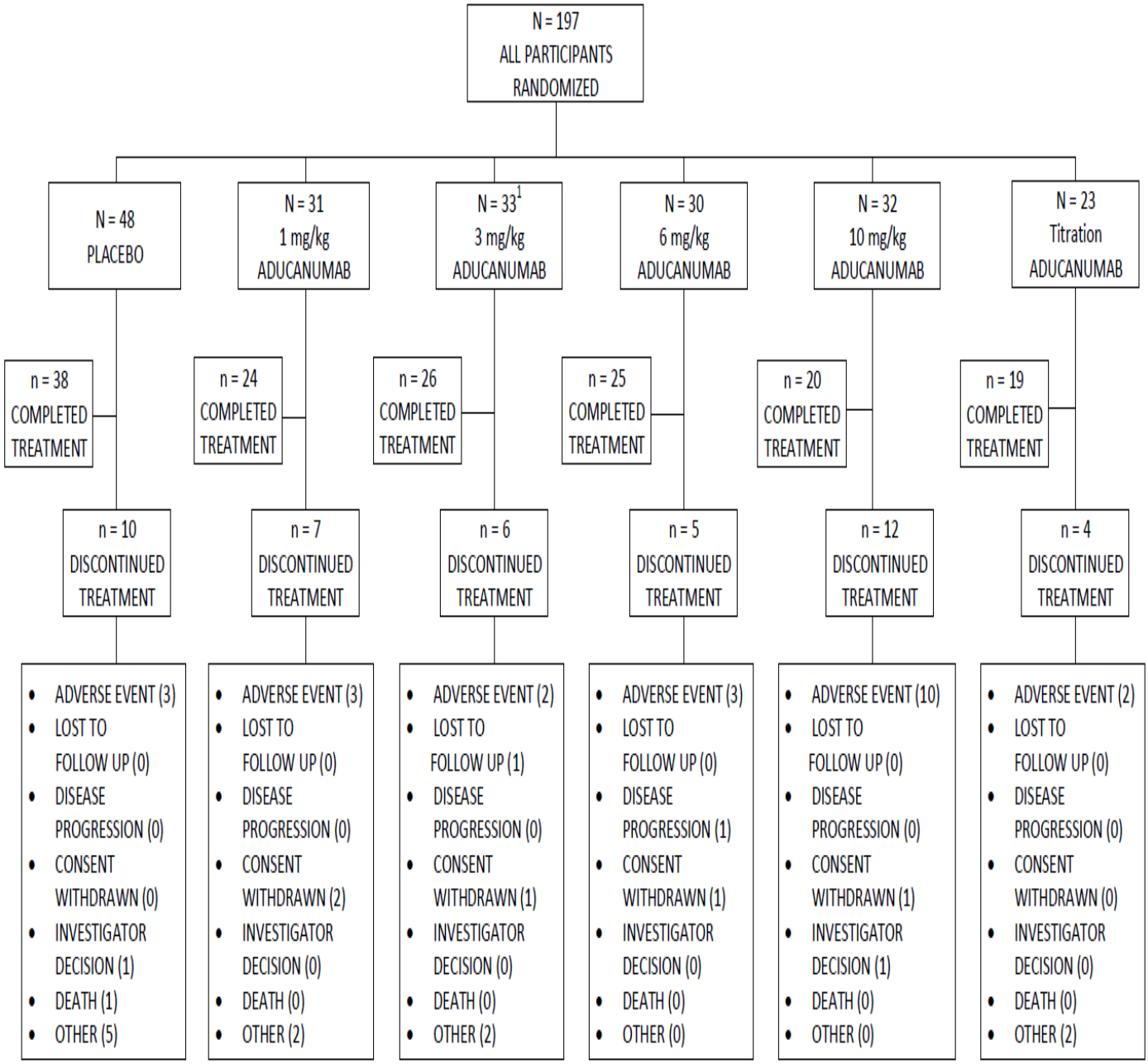
Sample Size Considerations

- In Arms 1-3, approximately 80 subjects were to be randomized 3:3:2 to aducanumab 1 mg/kg, 3 mg/kg, or placebo. In Arms 4-5, approximately 40 subjects were to be randomized 3:1 to 1 aducanumab treatment arm (10 mg/kg) or placebo. In Arms 6-7, approximately 40 subjects were to be randomized 3:1 to 1 aducanumab treatment arm (6 mg/kg) or placebo. In Arms 8-9, approximately 28 subjects were to be randomized 3:1 to 1 aducanumab treatment arm (1 mg/kg for the first 2 doses, 3 mg/kg for the next 4 doses, 6 mg/kg for the next 5 doses, and 10 mg/kg thereafter) or placebo. Combining all 9 arms, there were to be approximately 188 subjects of which approximately 30 subjects were to be in each treatment group (1, 3, 6, and 10 mg/kg), approximately 21 subjects in the titration group (up to 10 mg/kg), and approximately 47 subjects in the placebo treatment groups.
- The primary PD endpoint is change from baseline to Week 26 in ^{18}F-AV-45 PET signal in certain brain areas. A sample size of 30 subjects per treatment group would provide over 90% power to detect a treatment difference of 1 standard deviation with respect to the reduction of amyloid from baseline, based on comparison of each aducanumab group with placebo, at a two-sided significance level of 0.05, and assuming a dropout rate of 20%. In addition, under the same assumptions, a sample of 21 subjects per treatment group would provide over 80% power. Due to the exploratory nature of this trial, no formal adjustment for multiplicity was to be performed.
- Whole blood samples were to be obtained for ApoE genotyping at Screening. Subject enrollment was to be monitored (Arms 1-7) so that the ratio of ApoE4 carriers to non-carriers would be no more than 2:1 and no less than 1:2. If the treatment effect differs according to ApoE4 status, this sample size would provide at least 74% power within each ApoE4 stratum at a one-sided significance level of 0.1.

3.2.3.1.1 Sponsor's Results

Figure 28 shows patient disposition in study 103 with all placebo arms grouped together.

Figure 28 Study 103 Patient Disposition



[1] 1 participant did not receive any doses of study treatment.

Note: figure copied from page 65 of sponsor's study report

The sponsor submitted the interim analysis 2 (IA-2) results of Study 103 for FDA breakthrough designation.

Table 15 shows the sponsor's final analysis of study 103, based on a by-Visit ANCOVA model.

Reviewer's Comment: This is a completer's analysis which would not be adequate for a primary analysis when dropouts are not unexpected as the exclusion of dropouts may lead to bias.

Table 15 Sponsor's Completer's Analysis of Study 103

	Placebo	BIIB037 1 mg/kg	BIIB037 3 mg/kg	BIIB037 6 mg/kg	BIIB037 10 mg/kg	BIIB037 titration
Change from baseline at Week 54						
n	39	23	27	26	23	21
Adjusted mean	1.89	1.69	1.33	1.09	0.63	0.70
Standard error	0.350	0.441	0.413	0.423	0.446	0.499
95% CI	(1.198, 2.581)	(0.819, 2.564)	(0.517, 2.149)	(0.258, 1.930)	(-0.251, 1.511)	(-0.287, 1.686)
Difference with placebo		-0.20	-0.56	-0.80	-1.26	-1.19
95% CI for difference		(-1.308, 0.912)	(-1.612, 0.499)	(-1.855, 0.264)	(-2.356, -0.163)	(-2.343, -0.037)
p-value (comp. to placebo)		0.7249	0.2995	0.1398	0.0246	0.0432

NOTE: Adjusted mean for each treatment group, difference with placebo, 95% confidence interval and p-value were based on ANCOVA model at each timepoint. ANCOVA model was fitted with change from baseline as dependent variable, and with categorical treatment, baseline value and laboratory ApoE status (carrier and non-carrier) as independent variables.

(a) The efficacy analysis population is defined as all subjects who were randomized, received at least one dose of study treatment, and have both baseline and at least one post-baseline questionnaire assessment.

(b) Baseline population includes subjects in efficacy analysis population who have both baseline and at least one post-baseline assessment for CDR sum of boxes.

Note: this table was copied from page 110 of the sponsor's study report

3.2.3.1.2 Reviewer's Results

There was some potential for operational bias in this study due to having up to seven interim analyses with some sponsor personnel unblinded while the study was ongoing. Cases of ARIA were managed by the Biogen and the investigators.

The sponsor submitted a formal FDA breakthrough therapy designation request on 23 January 2015, based on interim analysis 2 of study 103 which included week 54 data for Arms 1-5 only. At this time the 10 mg/kg group had completed to Week 54 but the placebo groups for arms 7 and 9 that were utilized in the final analysis of study 103 were ongoing (arm 8 [titration to 10 mg/kg] and arm 9, the corresponding placebo, were just added to the design in July 2014 the same month that the last 10 mg/kg patient completed Week 54). Therefore, there are multiplicity and non-concurrent study arm issues with the pooled placebo group used in the final analysis.

The staggered arm design means that there is no direct dedicated randomization to validate dose response analysis (comparison of doses) or ITT interpretation of the comparison of individual doses to pooled placebo (placebo arms 3,7, and 9 had no chance of being randomized to 10 mg/kg and placebo arms 3 and 9 were not concurrent with 10 mg/kg [arm 4]). None of the direct comparisons directly supported by the staggered randomization(s) are nominally significant. This is a serious limitation as compared to a typical confirmatory phase 3 design.

Only the Arm 5 placebo is concurrent with the 10 mg/kg group. The estimated difference based on only these two groups was -1.13 [S.E.=0.70], p = 0.1178. For the titration to 10 mg/kg group the difference against the same cohort placebo was -0.57 [S.E.=1.08], p = 0.6046.
Table 16 compares the sponsor's non-randomization supported analyses and the randomization backed analyses.

Table 16 Study 103 Randomization Supported and Sponsor's non fully randomization supported Analyses

Comparison	LS Mean Treatment Difference	Standard Error of Difference	Nominal p-value
10 mg/kg vs same cohort placebo (randomization supported)	-1.13	0.70	0.1178
10 mg/kg vs pooled placebo (not randomization supported)	-1.08	0.54	0.0462*
Titration to 10 mg/kg arm vs. same cohort placebo (randomization supported)	-0.57	1.08	0.6046
Titration to 10 mg/kg arm vs. pooled placebo (not randomization supported)	-0.73	0.57	0.2044

*loses significance when post randomization starting of approved AD medications data is censored: p=0.09

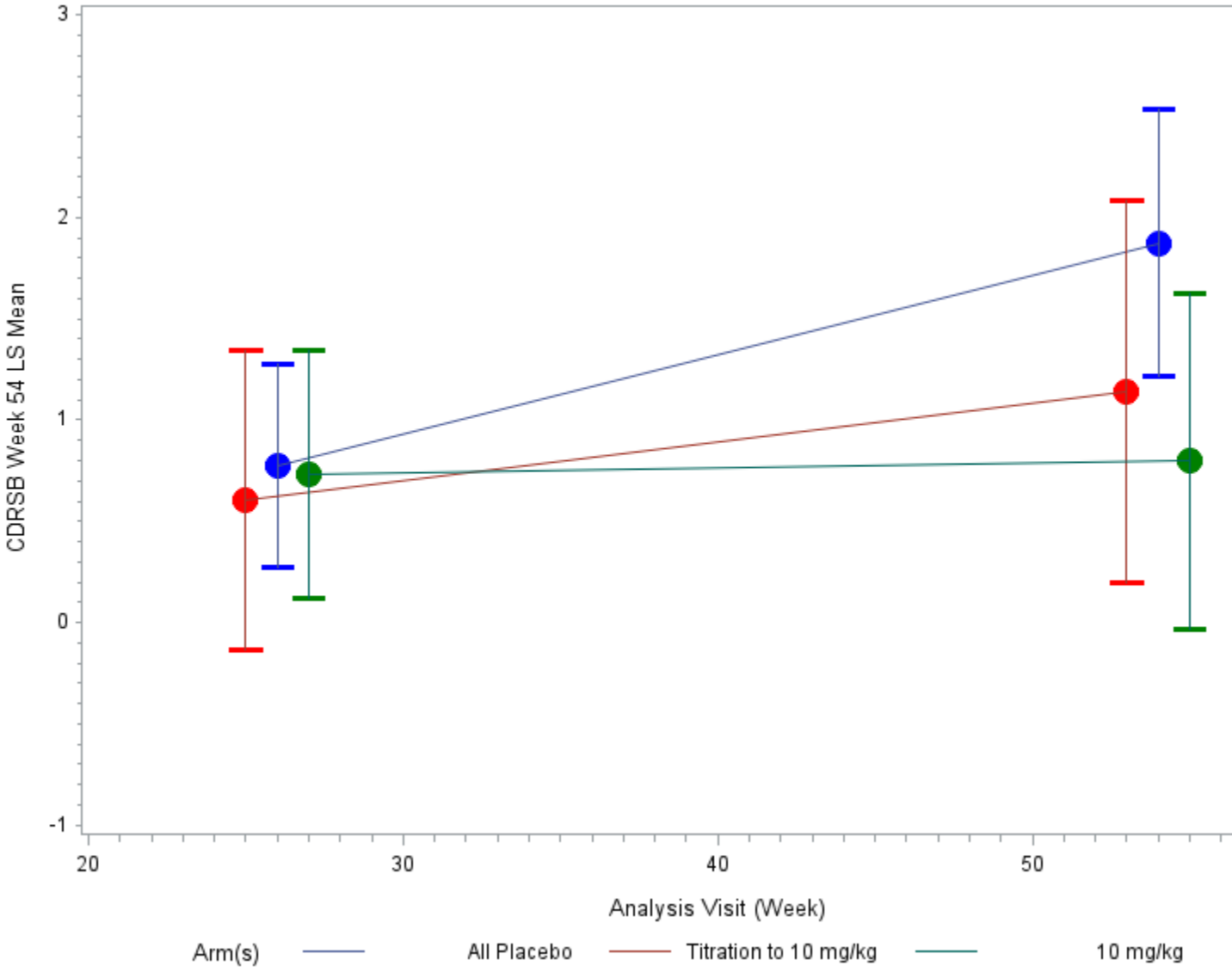
In the final study report the sponsor presents the analysis of Week 54 from an ANCOVA of Week 54 data only instead of the more appropriate MMRM (given the presence of dropouts 11% placebo [10% concurrent placebo] and 18% in 10 mg/kg). The MMRM was resigned to a sensitivity analysis by the sponsor. There were 20 randomized patients excluded from the ANCOVA.

The sponsor's ANCOVA result for 10 mg vs. **pooled** placebo was 1.26 [S.E.=0.556], p=0.0246. The more appropriate MMRM has a less favorable p-value for 10 mg vs. pooled placebo: 1.08 [S.E.=0.537], p=0.0462. Regardless, the comparisons with pooled placebo are nonrandomized comparisons. The pooled placebo LSMean at Week 54 was 1.90 [S.E.=0.34] worse than either study 301 or 302 in a shorter study and the Week 54 CDRSB variance was 4.57.

The MMRM estimate of the titration group vs. pooled placebo comparison is not significant.: 0.73 [S.E.=0.57], p=0.2044 as compared to the ANCOVA 1.19 [S.E.=0.58], p=0.0432. Dropouts were 5/44 for placebo overall and 1/22 for the titration group. The one BIIB037 titration group dropout had a 6.5 point increase from baseline to Week 26 and then dropped out. The by visit ANCOVA seems to be biased by informative censoring of this Week 26 CDRSB value: the week 54 LSMean for the titration group based on ANCOVA is 0.70 as compared to 1.14 for MMRM.

Figure 29 shows that neither the 10 mg/kg or the titration to 10 mg/kg had an effect at Week 26 on CDRSB. It shows the mean changes (and 95% confidence interval around the mean) in CDRSB by Study Visit as determined from MMRM for pooled placebo (dark blue), titration to 10 mg/kg (arm 8, red color) and 10 mg/kg without any titration (arm 4, green color). The means are connected for each group for visual aid, although it should be noted that the true progression curve between visits is unknown and may not follow the line. Note that the Visits were at Week 26 and 54 for all groups, but some separation was introduced in the figure to avoid overlapping.

Figure 29 Study 103 CDRSB profile for 10 mg/kg groups and Overall Pooled placebo



Three of the placebo (a fourth started right after) and three of the BIIB037 10 mg/kg had started other concomitant AD medications after baseline and prior to the Week 54 assessment. If these subjects are excluded from the analysis the difference on CDRSB is (based on 167 patients and 311 records) 0.896 [S.E.=0.534], p= 0.0954, not nominally significant. For the titration group, the result after excluding data from post-baseline started AD medications is 0.442 [S.E.=0.561], p=0.432.

The sponsor's MMRM result is also moderately influenced by a single outlier patient in the 10 mg/kg group, exploratory exclusion of this outlier patient results in a difference of .89 [SE=0.53], p=0.10.

In the sponsor's 10 mg/kg vs. pooled placebo comparison, more of the effect was in the APOE- subgroup in study 103 which is the opposite of what was seen in study 302. There was no significant interaction between treatment and APOE but this would likely be underpowered (interaction test for high dose vs. placebo at Week 54 by APOE, p=0.3746). The estimated Week 54 differences between 10 mg/kg and pooled placebo for APOE+ and APOE- subgroups were as follows.

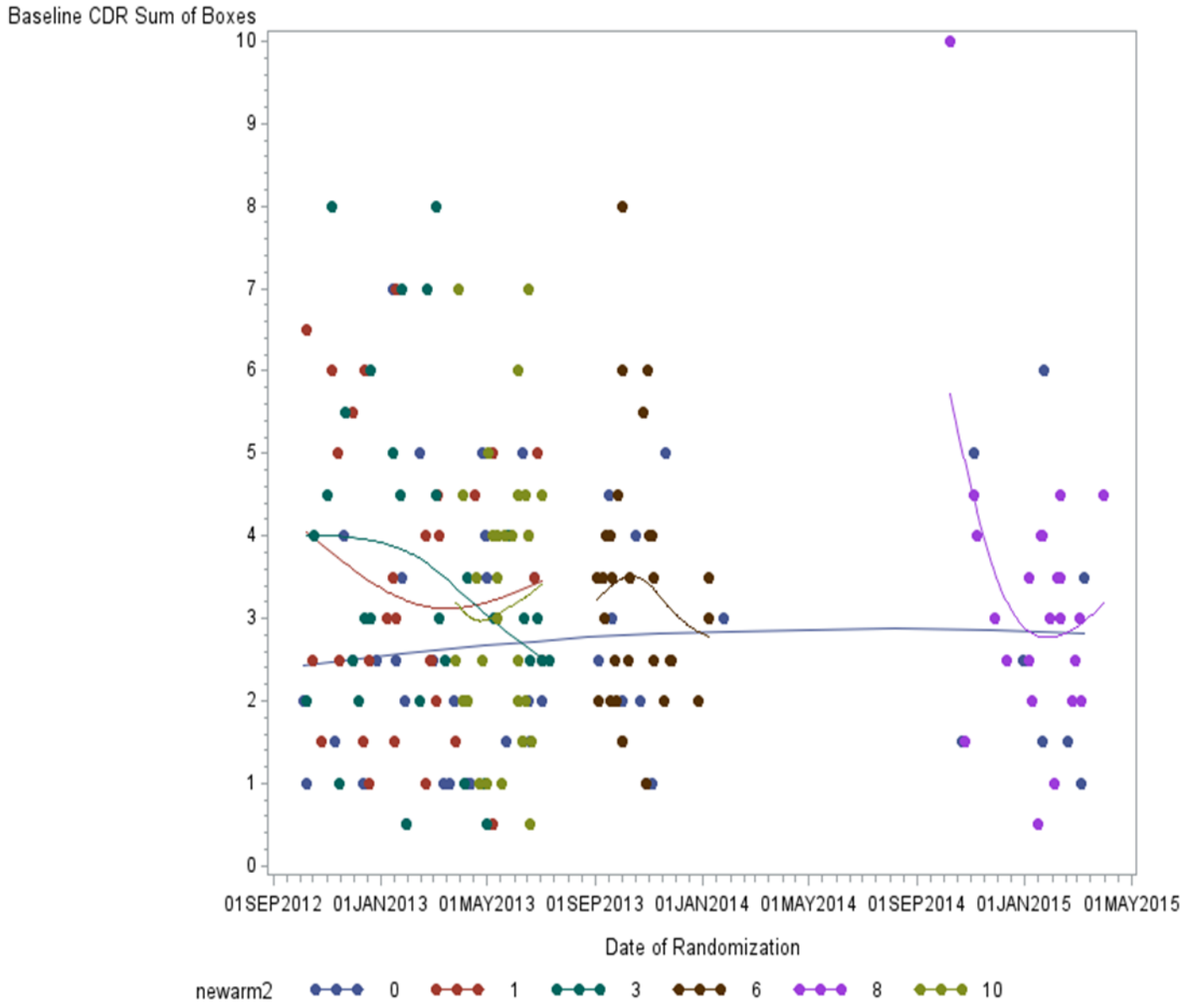
APOE+ 0.88 [S.E.=.642] p=.1716
APOE- 1.63 [S.E.=.956] p=.0906

.

There were also influential sites for study 103: without site 228 the 10 mg/kg estimated difference at Week 54 is 1.07 [S.E.=.55], p= 0.0561
Without site 219 the estimated effect is 0.95 [S.E.=0.53], p=0.0789.
Sites 229,236, and 201 were the most influential on the 10 mg/kg results. Exploratory exclusion individually led to p-values $\geq$ 0.09.

Figure 30 illustrates that all of the placebo was not randomized concurrently with the 10 mg/kg dose due to the staggered group randomization design. LOESS or local regression curves are drawn to aid the eye in the assimilation of the average CDRSB on a local [narrow window of time] level. Due to the staggering of the arms in time, the comparison of 10 mg/kg with all placebos may be biased in case of imbalances and, regardless, lacks the support of direct randomization between both complete groups. In fact, there are imbalances between 10 mg/kg and pooled placebo. For example, overall pooled placebo was 56% female but 10 mg/kg was only 41% female and, overall, placebo was 75.0% APOE carriers as compared to 65.6 % for 10 mg/kg. Also, some sites had no randomized patients from one group or the other, i.e., either pooled placebo or 10 mg/kg. The overall comparison may be confounded with these differences. If the titration groups which add some extra APOE carriers for pooled placebo are excluded the 10 mg/kg vs all other placebo is not significant 1.05 [S.E.=0.57], p=0.0689 and the titration cohort was only added to the protocol when the 10 mg/kg cohort was completing the placebo controlled portion. Furthermore, as noted above, if only the concurrent placebo arm is used 10 mg/kg is not significant 1.13 [S.E.=0.70], p=0.118.
Figure 30 illustrates the staggered arm design while showing the baseline CDRSB by arm. All placebo arms are grouped together. It is clear that there are placebos (dark blue) randomized both well before and well after the 10 mg/kg arm (olive). The number of the "newarm2" arm in the legend refers to the dose, except that newarm2=8 refers to the titration to 10 mg/kg arm.

Figure 30 Baseline CDRSB by Arm showing Staggered Design



Note: 8=Titration to 10 mg/kg other numbers are the arm's mg/kg doses (without any titration) or 0 for corresponding placebos

3.2.4 Response to Sponsor's Rebuttal of Appendix 2 to Advisory Committee Briefing Package

Note that the appendix 2 of the briefing package was based on the June 2019 dataset, which was the data evaluated by the collaborative workstream, since the June 2019 dataset was the first unblinded "final" dataset and there was only a difference in total CDRSB record counts of 4 out of 3716 total CDRSB records between the June 2019 and final BLA datasets (July 2020).

The sponsor's simulated probability of study 302 being false positive is not appropriate and is post hoc. These simulations don't give the primary endpoint the due prominence that it was allocated. They treat all endpoints as equals which doesn't reflect the prespecification. We can show that the false positive probability or p-value as defined will decrease as more variables are added even when some effects on added variables are zero or in the wrong direction. For simplistic articulation, let's assume that the false positive rate for testing only one endpoint A = 0.012.

Now add endpoint B and assume the trial shows drug is essentially equal to placebo on endpoint B; that is, the false positive rate for only endpoint B is essentially 0.50.

Considering both A and B jointly, the false positive rate so defined by the sponsor is $0.006 (= 0.012 \times 0.50)$, if the two endpoints are uncorrelated. If the two endpoints are positively correlated, the false positive rate may be higher than 0.006, but still much smaller than 0.012 unless the two endpoints are perfectly correlated. We see that adding more variables even with no effect will only decrease this probability. So, post hoc adding of more variables can make it look more impressive than it really is. In fact, if in truth drug = placebo on both A and B, a false positive can occur from the outcome of endpoint A winning or the outcome of endpoint B winning; consequently, the false positive probability is at least 0.50 without considering due prominence between A and B.

We do not believe that the false positive calculations by the sponsor capture all false positives that are as extreme; this is more complicated in multiple dimensions than for the usual one dimension primary endpoint. The false positive probability is a probability of a false rejection region. The region is constructed with test statistic (or statistics) and observed value (or values). The sponsor's definition of the rejection region in the calculation is incorrect. Let Y_1, Y_2, Y_3, and Y_4 be the test statistics for CDRSB MMSE, ADAS-cog13, and ADCS-ADL-MCI, y_1, y_2, y_3, and y_4 be their observed values. For each endpoint, one rejects the null hypothesis when the test statistic is better than the observed value. With adjusting signs for some endpoints, $\{Y_i > y_i\}$ represented the rejection for endpoint i in the sponsor's calculation. The sponsor defined the rejection region as $\{Y_1 > y_1, Y_2 > y_2, Y_3 > y_3, Y_4 > y_4\}$, following similar thinking in the one-dimensional case. With the same logic, the no-rejection region would be $\{Y_1 \leq y_1, Y_2 \leq y_2, Y_3 \leq$

y_3, $Y_4 \leq y_4\}$. However, $P\{ Y_1 > y_1, Y_2 > y_2, Y_3 > y_3, Y_4 > y_4 \} + P\{ Y_1 \leq y_1, Y_2 \leq y_2, Y_3 \leq y_3, Y_4 \leq y_4\} < 1$. The false rejection regions like {at least one but not all $Y_i > y_i$}, 14 in total, were missing in the sponsor's calculation. Some of these regions include the significant outcomes of the primary endpoint, and they should be considered rejection regions. For the N-dimensional case, the total number of rejection regions (those numerically favoring drug) is 2^N-1. With the sponsor's definition of false positive probability, only one region would be used, and a total of $2^N -2$ regions would be left out. If we add the testing of the low dose in addition to testing of the high dose because it was part of the prespecified plan, then the number of false positive regions that should be included becomes $2^8 -1 = 255$.

For example if CDRSB, MMSE, and ADAS-cog13 differences were doubled to -.78, 1.2, and -2.8, respectively, but ADCS-ADL-MCI was .01 worse, 1.69, then over all 4 hypotheses this is more extreme, further from the null statistically (in standardized Euclidean distance), but this case would not be counted by the sponsor's criterion, just because ADCS-ADL is .001 worse, but the others compensate by being much better, so statistically it should be counted. This is just one isolated point but volumes in the sample space with actual probability mass can be similarly defined that are not captured by the sponsor's criterion and yet are statistically more extreme than the observed.

We control the probability of at least one null hypothesis being falsely rejected at .05 two-sided, so ignoring the low dose outcome on the primary endpoint is not adequate. When the prespecified testing strategy is violated like this, the type I error probability can be much higher. For example, for study 302, an equally impressive post hoc analysis can be constructed such that the false positive probability including all 8 hypotheses (for 2 doses and 4 endpoints) is at least 0.7578 (this is the p-value of MMSE for low dose), if the prespecified due prominence for the high dose primary endpoint and the pre-specified testing hierarchy are ignored. The Sponsor acknowledges that Type I error is between .09 and .10 considering key secondary endpoints on page 36 of their response to Appendix 2 of the advisory committee briefing package.

We judge significance of studies on the basis of the prespecified primary endpoint, not the key secondaries. The primary endpoint showed an increased placebo effect after PV4, the mid study protocol amendment, in both studies. Since the high dose did not improve relative to the low dose, and in fact was numerically worse than low dose after the amendment in study 302, i.e. when the high dose dosing was highest, it seems more likely that the post-PV4 improvement relative to placebo is more related to placebo worsening than to dose increase for the high dose. The apparent improvement of the high dose in 301 post-PV4 shows the same phenomenon. Although the high dose is numerically better than low dose in post-PV4 by a very small amount,.07, on the primary endpoint (even less in the APOE carriers who got the dose increase), it is essentially the same as the low dose and placebo worsening post-PV4 was even more dramatic in study 301. Therefore, the apparent improvement of the high dose post-PV4 could be more due to the simultaneous placebo worsening than the dose increase. It's true that a

significant difference from placebo in the prespecified analysis is enough for a study. The problem here is the substantial evidence question, that 301 failed and, in fact, even had placebo numerically better than the high dose overall. Furthermore, the mid-study amendment complicates the design of the trial and the interpretation of the data.

Although it might not be zero, there is at best a weak relationship / correlation between change in CDRSB at Week78 and change in primary composite SUVR at Week 78 in 301 and 302 whether based on the group level difference from placebo or at the patient level. Change in SUVR on brain imaging was only collected in a convenience sample of about 1/3 of patients in the phase 3 trials, so it is exploratory. One key difference in the statistical review and sponsor clinical endpoint vs SUVR imaging correlation analyses is the summary level; that is, one is the patient level correlation while the other is the group mean level correlation. The group level correlation assumes equal placebo progression across studies but they are considerably variable, and study 103 was about 1/3 shorter but still had a faster placebo progression (1.90 vs. 1.75 and 1.54) and has much wider confidence intervals due to a much smaller size. If we plot the group level relationship between CDRSB and SUVR without subtracting placebo, then 10 mg/kg in 103, the 103 high dose equivalent, appears to be more of an outlier than the study 301 high dose group. The group level correlation plots did not present the correlation or the uncertainty of the group means. Regardless, the magnitudes of both the group level and patient level correlations are small, .18 and .14. Furthermore, per the sponsor's calculation on page 125 of their ise-appendix-g6 document, the proportion of the Week 78 clinical treatment effect in CDRSB explained by Week 78 change in SUVR was only 33% for the high dose and the corresponding 95% confidence interval did not exclude 0% explained. Thus, there is no compelling evidence that Week 78 change in SUVR is a surrogate and it is not clear that Week 78 change in SUVR predicts change in Week 78 CDRSB in a meaningful way.

We disagree with the sponsor's assertion that the reasons for failure of 301 are sufficiently well understood, since their reasons are post hoc (e.g. "rapid progressors") and depend on non-randomized actual dosing subsets and there exist counter arguments to their assertions (e.g., why is the low dose better than high in 302 after the PV4 high dose increase?).

Although in some sense consistent, the low dose effect was not nominally significant in either study and the consistency is diminished by multiplicity, it must be judged under the hierarchical plan for type I error control. The ADAS-Cog and ADCS-ADL were designated 3rd and 4th in the hierarchy, they can't be elevated after the fact and the trends must also be judged in the context of the hierarchical plan. Taking these out of the proper context would increase the risk of type I error. Being numerically worse on the prespecified primary endpoint in 301 should carry more weight than non-significant trends on the low dose and the 3rd and 4th endpoints.

Missing data was very high proportion in the ITT dataset. Missing data may not be MAR due to changes in enrollment characteristics later in the study. MMRM implicitly imputes missing data

from completers but there may not be enough overlap in model covariates between dropouts and completers due to changes in enrollment, especially with 70% missing. The implicit imputation may be slightly off due to omission of time dependence of model effects that nonetheless appear significant. Randomization helps but such a high amount missing may overwhelm this protection. We have 3 issues: very high missing, some differences in characteristics between completers and dropouts in model effects impacting outcome, and potential model bias due to not incorporating the time dependence of these effects which the data suggests is important, This is a recipe for trouble and we are on thinner ice with differences between dropouts and completers in country for example which the data suggests is an effect modifier. We agree that the results are reasonably consistent between the ITT and OTC but the OTC is more reliable due to a more reasonable amount of missing data and it has a slightly higher p-value, i.e, farther from the level considered for an isolated single study approval.

The sponsor exaggerates the consistency of subgroup effects in study 302. All of the 80 subgroup adjusted results are correlated because they are estimating the same overall effect. It is misleading to treat them as if they are independent. The APOE non-carriers had smaller effects at week 78 for the high dose despite higher average dosing than APOE carriers, and this was consistent across all 4 key endpoints and for both dose vs. placebo comparisons. In fact, there was a significant interaction for the first key secondary endpoint, p=.0096.

It's true there wasn’t more ARIA after the amendment and that there was some in placebo, but there was still much more in drug groups and excluding post-ARIA data is not conclusive because it is post-hoc and leads to a non-random imbalance in follow-up.

The sponsor had originally planned for the possibility of one interim analysis for efficacy using an O’Brien Fleming efficacy stopping boundary. The final analysis at about 60% complete Week 78 information relative to the planned total would not quite exceed that boundary if the boundary was applied. This case is unique in the sense that the futility stopping did not offer the chance for the trials to continue to the planned trial end. It is still arguable whether using the full alpha is appropriate; this issue is another less than ideal characteristic of this application.

The principal component analysis presented in this reviewer’s advisory committee backup slides may have been inappropriate (though it is mathematically accurate), but it doesn’t change the overall conclusions. The sponsor’s principal component analysis also seems suboptimal with it’s assumption of normality for discrete valued individual items (also, less likely to span the range of item scores and/or be normally distributed in the early stages of disease).

3.3 Evaluation of Safety

Incidence of falling adverse events was 3.2 % higher in the high dose than placebo (Table 17 showing common adverse events).

Table 17. Adverse Events With at Least 5% Incidence in BIIB037 10 mg/kg and 2% Higher Incidence Than Placebo –Pool A1

	Placebo (N=1087) n (%)	BIIB037 3 mg/kg (N=760) n (%)	BIIB037 6 mg/kg (N=405) n (%)	BIIB037 10 mg/kg (N=1033) n (%)	BIIB037 total (N=2198) n (%)
Number of subjects with any event	945 (86.9)	700 (92.1)	347 (85.7)	946 (91.6)	1993 (90.7)
Amyloid related imaging abnormality-oedema/effusion	29 (2.7)	223 (29.3)	83 (20.5)	362 (35.0)	668 (30.4)
Headache	165 (15.2)	161 (21.2)	58 (14.3)	212 (20.5)	431 (19.6)
Amyloid related imaging abnormality-microhaemorrhages and haemosiderin deposits	71 (6.5)	141 (18.6)	50 (12.3)	197 (19.1)	388 (17.7)
Fall	128 (11.8)	105 (13.8)	50 (12.3)	155 (15.0)	310 (14.1)
Superficial siderosis of central nervous system	24 (2.2)	91 (12.0)	23 (5.7)	151 (14.6)	265 (12.1)
Diarrhoea	74 (6.8)	62 (8.2)	27 (6.7)	92 (8.9)	181 (8.2)

NOTE 1: A subject was counted only once within each preferred term (MedDRA version 22.0).
NOTE 2: Preferred terms are presented in decreasing frequency of the table's BIIB037 10 mg/kg column.
NOTE 3: Preferred terms are displayed if the incidence is at least 5% in the BIIB037 10 mg/kg and the incidence difference compared to placebo is at least 2%.

Note: Table copied from page 64 of Sponsor's Clinical overview document

This was also statistically significant when analyzed by time to first fall (Hazard Ratio= 1.33 [1.05, 1.68] .p=0.0166) or number of falls adjusted by time at risk, risk ratio =1.30 (1.12,1.52)

The recent International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) addendum R1 on statistical efficacy ICH E9 (R1) stressed the importance of accounting for potential biasing intercurrent events such as death in primary statistical analyses.

Therefore, if a Joint rank of time to first fall and cognition/function is conducted as a benefit risk analysis then the high vs. placebo comparison in the OTC subset of Study 302 gives an estimated difference in ranks of: 36.6 [S.E.=28.8] p=0.2040. Note that this analysis method imposes a heavy penalty for falls, giving patients with them a rank worse than any observed Week 78 CDRSB in patients without any falls. Two patients who both had falls are ranked by the time to first fall and two patients without falls are ranked on the CDRSB changes at Week 78. The resulting (joint) rank sums are analyzed adjusting for the same covariates the sponsor specified in the primary analysis model except for those involving Visit because only the last common Visit between each pair of patients is used to determine the rank for the pair.

There were 6 high dose and 5 placebo deaths in the OTC or Died population (N=643 patients from placebo or high dose) of study 302. A joint rank analysis of CDRSB change and survival gave an estimated Week 78 difference in rank sums of 64.05 [S.E.=28.63], p=0.0256. Thus, a joint rank of CDRSB and survival has the same conclusion as the primary analysis (ignoring survival), since deaths were rare in the trial.

Please see the Clinical safety review for a review of general safety.

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race, Age, and Geographic Region

4.1.1 Gender, Race, and Age

SEX

In Study 302, 51% and in Study 301 52% were Female. In study 302, both high and low dose differences from placebo are smaller for females and the SEX*VISIT interaction in 302 is marginally significant p=0.0974 suggesting a potentially different CDRSB trend over time by Sex. Table 18 shows estimated treatment effects (high minus placebo) for the high dose by Sex and by Study as well as pooled across phase 3 studies. Pooling over identically designed studies (while allowing for a study effect) is suggested in the integrated summary of effectiveness guidance for subgroup estimation because subgroups, tend to be underpowered and have higher variability than the overall ITT population in individual studies. Females numerically favored placebo in Study 301 and in Study 302 had an estimated effect less than ½ that of the estimated

effect in Males. In the pooled analysis the high dose effect on CDRSB at Week 78 in Females was very close to 0.

Table 18 Differences on CDRSB by Sex Subgroup

Study	Group	Est. Diff. from Placebo	Std. Error
301	Female hi78	+0.2423	0.2086
301	Male hi78	-0.2023	0.2177
302	Female hi78	-0.2018	0.2145
302	Male hi78	-0.5674	0.2215
301/302 Pooled	Female hi78	+0.0202	0.1496
301/302 Pooled	Male hi78	-0.3848	0.1553

AGE

The mean age in 302 is 70.7 (median 72) and 80% were ≥65 years of age. The SAP states that age subgroup categories would be Age < 65, 65-74 and ≥75. Proportions among all groups falling into these categories were .23, .47, and .30 in 301 and .20 , .47, and .33 in study 302. Three factor interaction (AgeGroup*Treatment*Visit) in 302 had a p-value of 0.1023 (was 0.0919 in the June 2019 dataset first shared with the Agency).

When age is treated as continuous rather than grouping into the three categories the p-value for the 3-way interaction between Age, Treatment Group, and Visit is 0.0068. Testing only the two and three way interaction effects gives a p-value of 0.0075 . The implication of the interaction is that the estimated high dose effect was mainly in the highest age group (the model suggests no difference at or below Age 61 and better for high dose by .35 points for every 10 years above that).

Table 19 shows estimated treatment effects for the high dose (high minus placebo) by Age Group (<65,65-74,≥75) on CDRSB at Week 78 and by Study as well as pooled studies. The <65 group was in the wrong direction (placebo was numerically better) in 301 and in Study 302 had an estimated effect less than almost ¼ the size of the ≥ 75 group effect, as did the 65-74 age group.

Table 19 CDRSB High Placebo Differences on CDRSB at Week 78 by Age Group and Study

Study	Group	Est. Diff. from Placebo	Std. Error
301	<65 hi78	+0.2965	0.3029
301	65-74 hi78	-0.05108	0.2218
301	≥ 75 hi78	-0.08155	0.2781
302	<65 hi78	-0.2215	0.3400
302	65-74 hi78	-0.1926	0.2218
302	≥ 75 hi78	-0.8054	0.2748
301/302 Pooled	<65 hi78	+0.03750	0.2277
301/302 Pooled	65-74 hi78	-0.1218	0.1568
301/302 Pooled	≥ 75 hi78	-0.4435	0.1955

Race

In Study 301, 75.2% were White, 10.6% were Asian, and 14.2 % were classified as Other (including Black and Indian and/ or not reported).

In Study 302, for Race 78% were White, 8% Asian, 14% were classified as Other (including Black and Indian).

Race was actually not reported for 12.7 % in 302 and 13 % in 301.

Table 20 shows estimated high dose treatment effects (high minus placebo) at Week 78 by Race for individual studies, as well as for pooled studies.
There was some suggestion in study 302 of possible differential profiles and treatment effects by Race (more so in the June 2019 dataset Race Main effect p=0.0235; Race by Visit interaction p=0.0774 and Race*Visit*Treatment 3 way interaction p=0.1670). In Study 302 the estimated high dose treatment difference vs. Placebo on CDRSB at Week 78 in Asians was more than 2.5

times larger than the effect in Whites or Others, although the standard error was also larger due to limited sample size in this subgroup.

Table 20 Pooled and By Study Analysis of Estimated High Dose Treatment Effects by Race Groups for CDRSB at Week 78

Study	Group	Est. Diff. from Placebo	Std. Error
301	Asian hi78	0.08	0.51
301	White hi78	-0.16	0.17
301	Other hi78	1.00	0.40
302	Asian hi78	-1.08	0.68
302	White hi78	-0.39	0.17
302	Other hi78	-0.15	0.44
301/302 Pooled	Asian hi78	-0.50	0.43
301/302 Pooled	White hi78	-0.27	0.12
301/302 Pooled	Other hi78	0.43	0.30

4.1.2 Geographic Region

Enrollment by geographic region was not the same in studies 301 and 302 (Table 21). Several countries were only involved in one of the two studies (Australia, Austria, Great Britain, Denmark, Korea, Puerto Rico, Taiwan in 301; Belgium, Chechnya, Finland, Netherlands, Poland, Sweden in 302). The US proportion was slightly higher in 301: 46.3 vs. 39.8% in 302. There were 13 countries in study 302. The distribution of regional enrollment also changed over the course of the studies which may confound the impact of protocol amendment 4 (allowing the APOE+ high dose to reach 10 mg/kg instead of only 6 mg/kg in earlier protocols).In 302 Japan increased from 1.7 to 11.9% and Germany increased from 2.8 to 11.1%, while Poland decreased from 19.1 to 6% after PV4 and these countries were impactful on the high dose effect in the same direction as the enrollment change from before to after, thus making an assessment of the APOE+ dose increase from 6 mg/kg to 10 mg/kg in protocol amendment 4 confounded with Country enrollment changes from pre-PV4 to post-PV4.

Table 21 Enrollment by Country across Studies 301 and 302

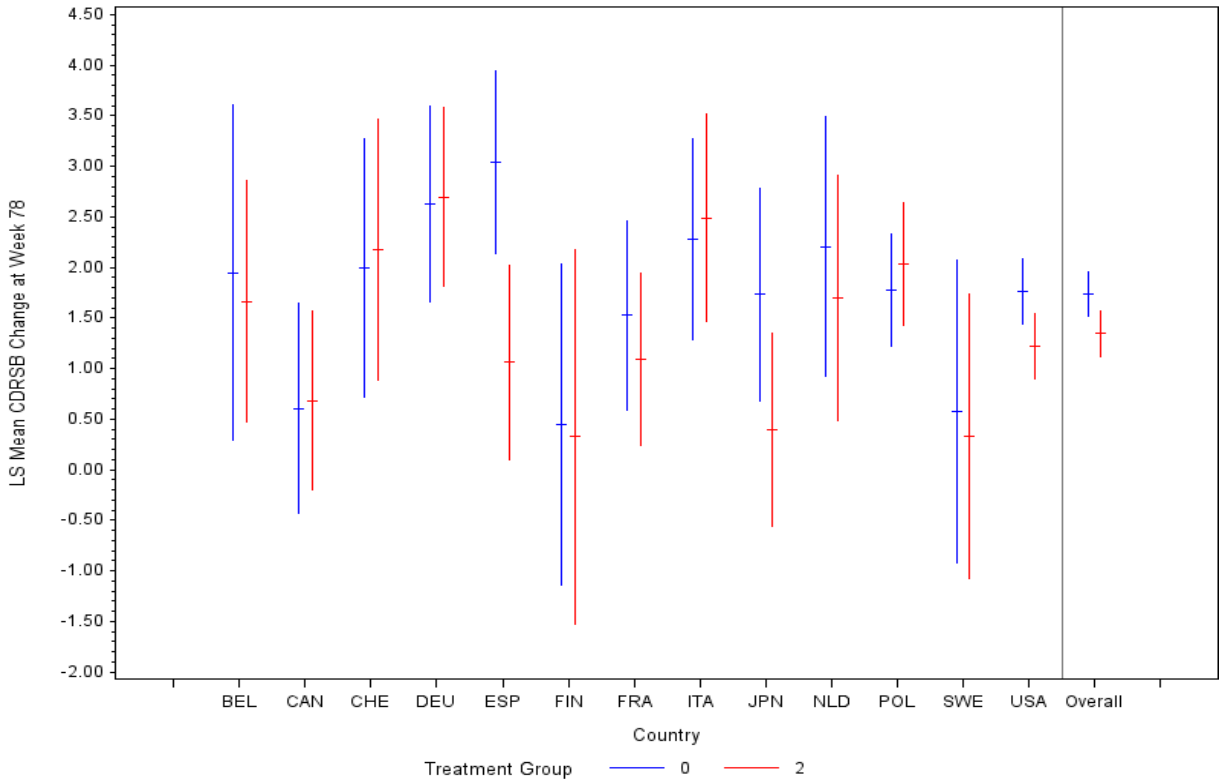
Country	Statistic	Study 301	Study 302	Overall
AUS	N	102	.	102
	Percent	6.2	.	3.1
AUT	N	9	.	9
	Percent	0.5	.	0.3
BEL	N	.	49	49
	Percent	.	3.0	1.5
CAN	N	83	96	179
	Percent	5.0	5.9	5.4
CHE	N	.	51	51
	Percent	.	3.1	1.6
DEU	N	96	122	218
	Percent	5.8	7.4	6.6
DNK	N	23	.	23
	Percent	1.4	.	0.7
ESP	N	105	78	183
	Percent	6.4	4.8	5.6
FIN	N	.	33	33
	Percent	.	2.0	1.0
FRA	N	60	78	138
	Percent	3.6	4.8	4.2
GBR	N	80	.	80
	Percent	4.9	.	2.4
ITA	N	116	81	197
	Percent	7.0	4.9	6.0
JPN	N	100	121	221
	Percent	6.1	7.4	6.7
KOR	N	50	.	50
	Percent	3.0	.	1.5
NLD	N	.	47	47
	Percent	.	2.9	1.4
POL	N	.	193	193
	Percent	.	11.8	5.9
PRT	N	47	.	47
	Percent	2.9	.	1.4
SWE	N	.	37	37
	Percent	.	2.3	1.1
TWN	N	13	.	13
	Percent	0.8	.	0.4
USA	N	763	652	1415
	Percent	46.3	39.8	43.1
All	N	1647.00	1638.00	3285.00

In study 302, three regions were prespecified for the Region effect in the primary analysis model: Europe/Canada, Asia, and the United States. There was a lot of variability within the Europe/Canada region's constituent countries' CDRSB outcomes. In four of the eleven

Europe/Canada region countries, LSMean Differences on CDRSB at Week 78 numerically favored Placebo compared to the High Dose. A test of whether including all of the countries fit the data better than just including the three regions was nominally significant, indicating that including the separate countries fit the data better(a Europe/Canada/Australia Region contrast has p=0.0004 against equality of this Region's constituent countries). This indicates that the primary analysis model's grouping of the European, Canada, and Australia constituent countries together into a 3 category region effect with 1 category for the 11 non-US, non Asian countries is inappropriate because there is significant variability among the Europe/Canada/Australia CDRSB change outcomes.

Figure 31 shows Placebo and High Dose group LS means with 95% confidence intervals for CDRSB at Week 78 by Country. The overall LS means are displayed in the far right section. The Europe/Canada/Australia region used in the analysis shows considerable variability (i.e., all lines except JPN 5th from the right and US 1st on the right).

Figure 31 Study 302 Placebo and High Dose LS means with 95% confidence intervals for CDRSB by Country



There was a significant country main effect (p<0.0001) on the Change from baseline in CDRSB when country was added to the primary model to check for consistency across countries. A likelihood ratio test of the primary analysis model augmented with country effects, country by

visit, country by treatment, and country by treatment by visit effects versus the primary model which did not adjust for any of these yielded a Chi square p-value of 0.0122, suggesting that there was statistically significant variation (lack of consistency) in the treatment effect across countries (Country Main effect p<0.0001; COUNTRY*TR01PG1N p=0.7644 ; COUNTRY*AVISIT p=0.0028 ; COUNTR*AVISIT*TR01PGN p=0.0095 note: TR01PG1N denotes the treatment group variable and AVISIT denotes the value of the analysis Visit variable [Week 26,50, or 78]).

An F test of just the last 3 terms representing different profiles by Country and Treatment Group (96 num DF) has a p-value of 0.0024. A joint test of Country*TR01PG1Nand country*AVISIT*TR01PG1N terms only, given country and country*AVISIT terms, has a p-value of 0.0543. Therefore, considering that the study is not powered for tests of interaction to say that the three way interaction is driven by the COUNTRY*AVISIT part is an oversimplification. On the contrary, there seems to be rather compelling evidence of inconsistency in treatment differences across countries.

An exploratory analysis excluding the US (which accounts for 40% of the overall population), was not nominally significant (Non-US 302 N=955 patients Hi-Pl Wk78 diff=-0.28 SE=0.22, p=0.1971). One can see in Figure 31 above, that Spain (ESP N=78) had the most rapidly progressed placebo response (largest mean and most favorable comparator for drug) among the countries. Exploratory exclusion of Spain (ESP) alone also resulted in a loss of significance for the Week 78 high dose difference from placebo on CDRSB: -.30 SE=.16, p=0.060 (Total remaining patients N=1504).

For Placebo group data alone Country by Visit interaction has a type 3 F test p-value of 0.0098. For the high dose the corresponding test has a p-value of 0.1785 and for the low dose 0.0034. Thus, there seems to be evidence of significant variation in CDRSB profile over visits between countries, particularly for placebo and low dose groups.

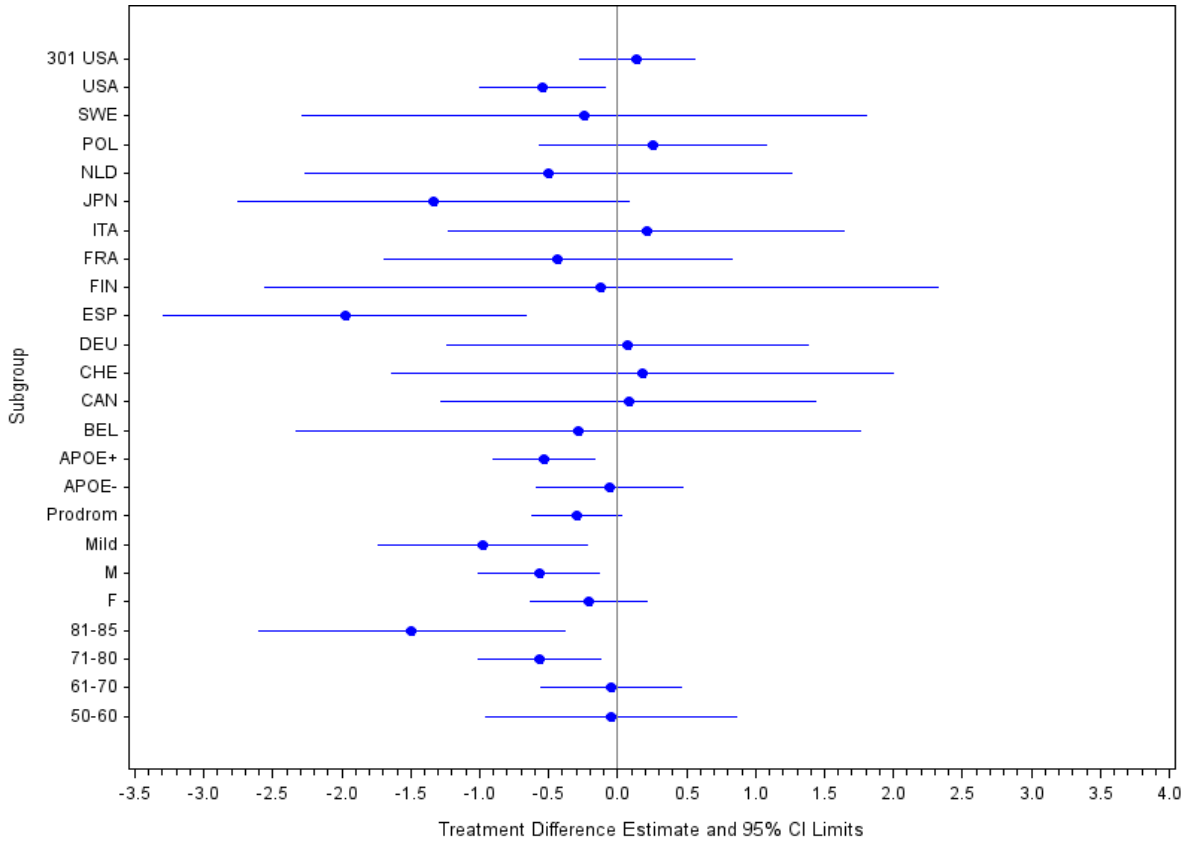
For 301 the corresponding tests of country effects and interactions are as follows.

Effect	Num. DF	Den. DF	F-statistic	p-value
COUNTRY	13	1467	4.37	<.0001
COUNTRY*TR01PG1N	26	1470	1.18	0.2410
COUNTRY*AVISITN	26	1904	1.50	0.0497
COUNTR*AVISIT*TR01PG	51	2007	1.10	0.2866

A joint F test of just the last 3 terms has a p-value of 0.0699. Therefore, at the least, 301 also suggests that CDRSB progression profiles varied significantly by country after adjusting for treatment groups and other primary model effects.

A forest plot of high dose treatment effects on CDRSB at Week 78 by Country follows in Figure 32. Lower CDRSB scores are better and the difference is presented as Aducanumab 10 mg/kg - Placebo, so that negative differences favor Aducanumab. The forest plot shows inconsistencies of high dose treatment effect at Week 78 in study 302 for important subgroups such as Country, Mild AD vs. Prodromal AD (less effect in Prodromal [interaction p=0.09]), Age Group (more effect in older [interaction p=0.06]), APOE (more effect in + [interaction p=0.15]). These inconsistencies might be important if one was evaluating whether study 302 could stand on it's own, which however would introduce selection bias given the 301 result.

Figure 32 Forest Plot of Change in CDRSB at Week 78 by Country and Other Subgroups



There were 7 countries included in both 301 and 302. Altogether they accounted for about 78% of the patients in both studies (80% in 301 and 75% in 302). Therefore, 25% of 302 patients were from countries not involved in 301. The pooled results by country, after allowing for study differences, for the high dose difference in CDRSB at Week 78 for the countries involved in both studies are as follows in Table 22. In this table negative differences favor the high dose. As can be seen in the forest plot of Figure 32 above, the high dose was in the wrong direction in the US subgroup of Study 301. The high dose effect at Week 78 in Japan, the US, Canada, and Italy among others had different signs (favoring drug or favoring placebo numerically) between 301 and 302.

Table 22 CDRSB High vs. Placebo Results at Week 78 Averaged over 301 and 302 by Country

Country Sample Size at Week 26 and 78	Country/Visit	Estimated Difference	Std. Error
171/85	CAN 78	-0.63	0.49
200/87	DEU 78	0.55	0.46
177/106	ESP 78	-1.08	0.44
133/89	FRA 78	0.20	0.49
194/94	ITA 78	-0.33	0.47
214/60	JPN 78	-0.54	0.52
1348/ 898	USA 78	-0.19	0.16

4.1.2.1 Individual Sites

There were 180 sites among the 13 countries in study 302. Note that the randomization was stratified by site and APOE carrier status.

Table 23 shows the effect of exploratory exclusion of select influential sites on the primary result.

Table 23 Study 302 Hi vs. Placebo LSMean differences on CDRSB at Week 78 after excluding a site of interest

Site excluded	#Patients	#Records	Estimate	Stderr	p-value
None	**1581**	**3712**	**-0.390**	**0.155**	**0.0120**
- 872(USA)	**1573**	**3690**	**-0.3484**	**0.155**	**0.0247**
- 856(USA)	**1573**	**3690**	**-0.3528**	**0.155**	**0.0227**
- 849(USA)	**1554**	**3647**	**-0.3674**	**0.157**	**0.0196**
- 669 (ESP)	**1564**	**3680**	**-0.3682**	**0.156**	**0.0185**

4.2 Other Special/Subgroup Populations

The dose was weight adjusted and yet, still baseline weight was found to be a nominally significant predictor of change from baseline in CDRSB (p=0.0909 [0.0288 in June 2019 dataset]) which also varied significantly by Visit (WEIGHTBL*VISIT, interaction p=0.0113). The sign of the estimated effect suggests that higher weights tended to have better CDRSB scores. For example, those above the median weight of 71 had an estimated high dose effect of -0.55+/-0.22 (S.E.), whereas those below had an estimated high dose effect of -0.25 +/- 0.22 (S.E.).

Baseline Alzheimer's disease stages (Prodromal/MCI or Mild AD) are important subgroups of interest for medical practitioners and drug effects can potentially vary by disease stage. There was a bigger effect in the smaller Mild subpopulation (18% of all 302 patients) in 302 (Figure 32

above and Table 24 below). The difference between these subgroups for the high dose treatment effect at Week 78 would be significant at the .10 level (Mild vs. MCI 0.68 [S.E.=0.42], p=0.1052), which may not be an unreasonable significance level for an underpowered interaction test. The baseline disease stage main effect and various interactions with it were somewhat compelling given the lower power for subgroups (

EFFECT	Numerator DF	Denominator DF	F statistic	p-value
ADBL	1	1551	42.47	<.0001
ADBL*TR01PG1N	2	1528	2.08	0.1255
ADBL*AVISITN	2	1138	14.20	<.0001
ADBL*AVISITN*TR01PG1N	4	1341	1.89	0.1100).

The difference was more compelling when High or Low differences from placebo were considered p=0.0324, because they were both in the same direction and the low dose was numerically worse than placebo in the prodromal subgroup (Low prodromal +.13 S.E.=0.41, Low Mild -.30 S.E.=0.17).

Table 24 shows the high minus placebo Week 78 CDRSB results by baseline disease stage (Mild AD or MCI/Prodromal AD) for each individual study and Pooled (interacting baseline disease stage group, treatment group, and Visit effects with Study).

Table 24 Study 302 Estimated CDRSB Change High Dose Differences from Placebo at Week 78 by Baseline Stage Diagnosis

Study	Group	Est. Diff. from Placebo	Std. Error
301	Mild hi78	+0.27	0.36
301	MCI hi78	-0.02	0.16
302	Mild hi78	-0.97	0.39
302	MCI hi78	-0.29	0.17
301/302 pooled	Mild hi78	-0.35	0.26
301/302 pooled	MCI hi78	-0.16	0.12

APOE

The randomization was stratified by APOE and 67% were carriers (APOE+). The treatment difference on CDRSB for the high dose in the APOE- (non-carrier) subgroup, by which the randomization was stratified, was very consistent across studies 301 and 302, very small and not nominally significant (Table 25). As discussed above in 3.2.1.4.2 the smaller estimated effect for APOE non-carriers than APOE carriers on CDRSB in Study 302 at Week 78 was also observed for all of the other key secondary endpoints (including a significant interaction for MMSE : APOE*TRT interaction p=0.0096). The difference between APOE – vs APOE+ high dose differences from placebo has p=.1511 at Week 78 for CDRSB, p= 0.07 for both MMSE and ADCS-ADL, all of these also favoring carriers.

Table 25 CDRSB at Week 78 for High vs Placebo by APOE and Study and Pooled

Study	Group	Est. Diff. from Placebo	Std. Error
301	apoegr3n - hi78	-0.06	0.27
301	apoegr3n + hi78	0.07	0.18
302	apoegr3n - hi78	-0.06	0.27
302	apoegr3n + hi78	-0.52	0.19
301/302 pooled	apoegr3n - hi78	-0.06	0.19
301/302 pooled	apoegr3n + hi78	-0.23	0.13

Concomitant Use of AD Medications at Baseline

Concomitant use of AD medications at baseline was also a primary model adustment factor and thus defines important subgroups of interest. Just over half used concomitant AD medications at baseline (51%). High dose vs. placebo treatment group differences were relatively consistent across those using AD medications at baseline and those not using in study 302 (No -0.44 [S.E.=0.23], Yes -0.35 [S.E.=0.22]). However, the CDRSB profile across visits appeared to vary between them as indicated by a nominally significant baseline concomitant AD medication use*Visit interaction test (ADCMBLFL*AVISITN p=0.0006).

5 SUMMARY AND CONCLUSIONS

5.1 Statistical Issues

Study 302 didn't complete according to the protocol because it was terminated for futility (<20% chance of success in either 302 or 301 study for either dose) based on the prespecified pooled analysis of it and study 301 that was used to project the second half of the data in each study when both were 50% complete. The data cutoff date was 26 Dec 2018, and the public futility announcement date was March 21, 2019. According to closed executive session meeting minutes, because the analysis showed futility, the DMC was provided with the efficacy and safety data from both the ENGAGE and EMERGE studies [221AD301 and 221AD302]. Furthermore, the DMC discussed various aspects of the data and the analyses, agreeing that even with careful review of the data and analyses provided, that there was no evidence of movement that would have resulted in a non-futility conclusion. In the final open DMC session the DMC informed the Biogen attendees that there were no additional analyses, beyond the pre-specified analyses, requested to reach their recommendation regarding the trials. The DMC stated that it was their unanimous recommendation that the trials be halted. The DMC asked if the Biogen attendees would like to review with the DMC the prespecified outputs for the futility analysis. Biogen agreed, and the analyses were reviewed in a joint session between Biogen attendees and DMC members.

There was considerable unblinding to manage ARIA events during the course of the study (ARIA-E events: 35% high dose [45% in APOE+], 26% and 2% for low dose and placebo in study 302, page 177 of June 14 2019 Type C briefing package). The absence of a significant impact of data collected after ARIA events does not necessarily imply that there was no bias due to unblinding due to ARIA; there is limited power for detecting such a difference. Moreover, one can't conclusively rule out an impact of those experiencing ARIA on the result because it requires making a comparison based on differential exclusions between the randomized groups (preferential exclusion of drug patients and/or censoring of drug arm data) and the resultant groups without ARIA to be compared are no longer as randomized and/or have differential follow-up and selection bias due to conditioning on a post-randomization event.

There was a major protocol amendment in the middle of the study modifying dosing for the APOE+ stratum high dose group. Changing enrollment by country over time and differential efficacy by country and variations in placebo response over time make it virtually impossible to elicit the effect of increasing the dose in protocol amendment 4. If the sponsor had not started phase 3 prior to completion of the phase 1B study, phase 3 could have started with the APOE+

group at 10 mg/kg and we would have a cleaner study and a clearer picture of whether or not more 10 mg/kg doses is important.

We should keep in mind that prior to amendment 4 the high dose is not the same for APOE+ and APOE-, 6 mg/kg and 10 mg/kg respectively, and also the moderately common occurrence of ARIA limited the dosing (35% of high dose patients had dose modifications). This would seem to possibly require drawing back from a blanket missing at random assumption for imputation of missing data for the high dose and call into question an imputation model not accounting for these differences. In 302 the occurrence of ARIA in the high dose was 1.5 times higher for APOE+ vs. APOE- prior to amendment 4 and 2.2 times higher post PV4 (note that ARIA in non-carriers was slightly higher pre-PV4 compared to post-PV4). Limitation of dose titration in the high dose was 2.4 times higher for APOE+ prior to PV4 and 3.4 times higher post PV4. The APOE- stratum had more 10 mg/kg doses but was worse on average than APOE+ in 4 out of the 4 primary and key secondary endpoints (and the low dose shows the same pattern) which calls into question the sponsor's assertion about the importance of the number of 10 mg/kg doses . The high dose is even in the wrong direction compared to placebo at Week 78 for APOE non-carriers on the first key secondary endpoint, MMSE, in Study 302.

Note that 302 has significant interactions with treatment*visit*Agegroup and treatment*visit*Country, as well as visit*agegroup and visit*country. Study 301 also has a significant country*visit interaction. These, among other interactions such as baseline disease stage group found by the reviewer, suggest that the primary MMRM model may not be correct. Since demographic and disease characteristics changed somewhat after PV4, those missing Week 78 have different characteristics and if the model is not correct the model may be biased given the large amount of missing data for the ITT population. Different demographics of those without the opportunity to complete such as more from the Asian region and DEU (Germany), more mild baseline disease stage, more symptomatic AD medications at baseline, and increased age may cause the model to not be valid under MAR since the model does not account for interactions with these variables and Visit, but the data suggests they are significant (should be included). With more than 40% missing data and the uncertainty of the primary model's validity under MAR (not including interactions that the data suggest are significant) it seems that the Opportunity to Complete Population result is more reliable and more relevant (also considering that Week 78 is the only significant timepoint in Study 302).

Regional differences may offer an alternative explanation for the 301 and 302 study outcome differences (less US in 302: 39.8% vs. 46.3% in 301; Poland 11% in 302, 0% in 301; AUS 6% in 301, 0% in 302; and several other countries only in one study) since 302 showed variation in high dose treatment group difference from placebo on the primary endpoint at Week 78 by Country (a nominally significant Country by Visit by Treatment Group interaction). Please see Table 21 for Country enrollment by study details.

The effect of the protocol version 4 change of dose for APOE+ high dose group is confounded with the following population enrollment changes from pre to post-PV4:

- Decrease in US: 7.5% Placebo, 11.1% Low Dose and 10.9% High Dose drops in percentages in US region from pre-PV4 to post-PV4,
- Decrease in Poland: for Placebo from 20.1% pre-PV4 to 5.5% post-PV4 (Placebo -14.6%, Low -11.4%, High -11.9% drops post-PV4 note: Poland numerically favored Placebo over High dose) Poland High dose effect is in wrong direction (+0.26 [S.E.=.42])
- Increase in Japan: Placebo 1.7% pre-PV4 to 12.6% post-PV4; (Placebo 10.9%, Low 9.7%, High 10.6% increases post-PV4 note: Japan had a big effect for high dose (-1.35 [SE=.73]) and may drive 302 result)
- Increase in Germany (DEU): Placebo 2.6% pre-PV4 to 9.7% post-PV4 (Placebo 7.1%, Low 7.0%, High 8.7% increases post-PV4)
- Decrease in Age Group (61-70 years) Placebo group proportion in this age group pre-PV4 is 40.6% vs. post-PV4 is 30.7%
- Increase in Age Group (71-80) Placebo proportion in this age group pre-PV4 is 41.0% vs. post-PV4 is 50.2% Placebo also had a 9.4% increase in the age ≥75 category from pre-PV4 to post-PV4 compared to 1.3% for the high dose group.

These enrollment changes are confounded with pre-PV4 to post-PV4 dose changes and if responses differ by countries (as the data suggests) it would be almost impossible to balance actual dosing subgroups across countries using propensity scores to get good matching. Regardless post hoc matching to exploratory subgroups can never reach the gold standard of a randomized comparison. Poland which is in the wrong direction for the high dose effect in 302 at Week 78 had higher enrollment pre-PV4 as compared to post-PV4 and Japan which had a big positive effect for the high dose had enrollment increased post-PV4 as compared to pre-PV4. Therefore, the effect of the protocol amendment 4 dose increase for high dose group APOE carriers is confounded with these enrollment changes.

More than 45% are missing Week 78, the only timepoint that showed nominal significance on the primary endpoint (45% for high dose and 47% for placebo). Neither the primary nor any of the key secondary endpoints was significantly different from placebo at Week 50 for the high dose in study 302. With only a single positive timepoint, no other positive timepoint to confirm it, it is not established that the progression was slowed even in study 302. There are no compelling correlations within the high dose group between change in Aβ SUVR composite with reference in the cerebellum, the primary biomarker, at Week 78 and Change in CDRSB at Week 78. In fact, in study 302 this unadjusted correlation was in the wrong direction. Furthermore, unlike the Solanezumab program, for Aducanumab there is no delayed start design to potentially support a slowing of progression. Even if there was, it would likely be problematic due to the early futility stopping since there would be associated dropouts and disruptions.

There was a blinded sample size increase [Updated sample size (from 450 to 535 participants per treatment group) condu0cted in November 2017] prior to the unblinded interim futility analysis but the conditional power for study 302 high dose was still only 58.63% at the interim for CDRSB based on the study 302 interim effect size only (N=782 the effect size was -.28, SE=.19, p=0.138) . If the final 302 high dose effect on CDRSB was real, the chance of 301 succeeding given it's sample size would be 0.755 {the SE would be sqrt(3.547*((333+293)/333/293))= 0.151, so the chance of success would be 1-pnorm(1.96-.4/.151)= .755}. The chance of observing a result as unimpressive as observed for 301 would be Φ(-.03/.151-.4/.151)=0.002. The estimated difference in the patients not included in the interim N=835 is -0.42 +/- 0.32 (SE) for high dose and -0.61 +/- 0.33 (SE) for the low dose (the low dose is numerically better than high after the interim!). This seems to conflict with the sponsor's assertion that protocol version 4's allowing the high dose to increase from 6 mg/kg to 10 mg/kg in APOE carriers optimized the dose response. Placebo showed a marked worsening after the interim (N=799, 1.75 [S.E.=0.23] pbo; 1.13 [S.E.=0.23] hi; 1.33 [S.E.=0.22] lo) vs. (interim: 1.53 [S.E.=0.16] pbo; 1.25 [S.E.=0.16] hi; and 1.43 [S.E.=0.16] lo). The sponsor's explanation requires considering a post-randomization event defined subgroup which has no proper non-counterfactual control, ability to tolerate drug without ARIA, for statistical justification, whereas the higher placebo response explanation for the better effect after the interim or lack of improvement of high dose compared to low post-PV4 requires no breaking of the randomization.

The correlations between endpoints need to be considered for properly assessing the chance that all of the primary and key secondary endpoints achieve nominal significance. The correlations among the primary and key secondary endpoints at Week 78 are substantial (Table 26). All of the endpoints were positive in 302 and they all were negative in 301. The chance of them all being nominally significant is increased as the correlations among them increase. In study 301 the correlations among the primary and key secondary endpoints are very similar to those in study 302 all > .40 in absolute value.

Table 26 Study 302 Correlations at Week 78 between changes from baseline on Primary and Key secondary endpoints

Prob > |r| under H0: Rho=0 Number of Observations				
	CDRSB	**MMSE**	**ADAS-cog**	**ADCS-ADL-MCI**
CDRSB	1.000	-0.557	0.494	-0.639
		<.0001	<.0001	<.0001
	877	876	866	862
MMSE	-0.557	1.000	-0.583	0.443
	<.0001		<.0001	<.0001
	876	880	869	864
ADAS-cog	0.494	-0.583	1.000	-0.397
	<.0001	<.0001		<.0001
	866	869	869	857
ADCS-ADL-MCI	-0.639	0.443	-0.397	1.000
	<.0001	<.0001	<.0001	
	862	864	857	864

Aducanumab showed dose dependent changes in composite SUVR Aβ plaque reductions with reference to the cerebellum, at Week 78. The effect on composite SUVR change was larger in non-carriers of APOE than APOE carriers, yet APOE non-carriers showed less effect on CDRSB change from baseline at Week 78. Why don't high dose non-carriers show a clinical effect in Study 302 if they got 10 mg/kg earlier, have less ARIA dose reductions and if they showed a bigger effect on composite SUVR Aβ uptake? This seems to call into question whether Aβ PET

composite SUVR with cerebellum reference is a surrogate. The sponsor's exploratory mediation analysis also suggested that the Week 78 SUVR change explained more of the clinical effect for the low dose than for the high, 36% vs. 33% and the confidence intervals did not exclude 0%. In fact, within the high dose group, there is actually no compelling correlation between Week 78 change in composite SUVR Aβ and Week 78 change in CDRSB. This seems to call into question how the amyloid biomarker could support a disease slowing claim. Baseline disease severity group (Mild AD vs. Prodromal AD) also showed this disconnect in the relationship between SUVR and CDRSB change, i.e., more SUVR change for Prodromal but less clinical change for Prodromal.

The current efficacy standard in AD is placebo controlled comparisons from clinical trials. An argument based on exposure response is circular. It presupposes that the drug is effective and/or that the SUVR is correlated with clinical cognitive and functional change, neither of which has been consistently shown across the phase 3 clinical data. Patients who were most compliant or tolerated the drug best and achieved the highest exposure were not randomized to that outcome and so there is no corresponding control group that they can be compared to without possible confounding due to selection bias.

Rapid progressors may be part of the reality of Alzheimer's and after unblinding it is too late to address them in a completed randomized study. A highly effective drug would probably not be likely to fail because of rapid progressors in a large study. Study 302 could just as well be the outlier relative to the true proportion of outliers in the natural progression. In fact, the range of CDRSB changes in Study 301 at 18 months appears consistent with the ADNI study (adnimerge_May15.2014 data). There are slightly more outliers in the high dose in 301 but that is worrisome in itself since they are consistent with the ADNI data and so should again raise doubts about the representativeness of the 302 result. Furthermore, robust regression, techniques (M estimation, least trimmed squares, MM estimation, S estimation) designed to be resistant to and downweight outliers, applied to the 301 Week 78 data still suggested no effect of the high dose compared to placebo and that it was numerically worse than the low dose (vs. Pl +0.027 [S.E.= 0.125], p=0.8315; vs. Low +0.063 [S.E. 0.125]). Without the worst change of +13 in the high dose group the 301 high dose vs. placebo result from the primary analysis model is +.0267 [S.E.=.1495], p=0.8581 as compared to +.0316 [S.E.=.1498], p=0.8330 including it. This shows that Study 301 is a big study and one outlier patient has limited influence (excluding the worst 3 Week 78 changes for high dose the difference from placebo is still just -0.013 [SE=.147], p=0.9294). Totally excluding the worst high patient instead of just the Week 78 observation the result is +.0072 [S.E.=.1487], p=0.9615, still in the wrong direction. More than one outlier in the high dose seems to be more of a systemic problem and should be more worrisome and harder to discount after the fact.

The primary objective of Study 103 was to evaluate safety and tolerability of multiple doses of Aducanumab in Alzheimer's patients and it was exploratory and hypothesis generating for clinical efficacy. in our view. Study 103 is an outlier among the three available studies. Study 103 had a much larger effect (300%) by Week 54 than 302 had at Week 78 and 302 showed nothing significant at Week 50 (note that titration was different for 10 mg/kg between studies 103 and 302, but placebo decline may also differ between 54 and 78 week studies,103 has a much bigger placebo decline at Week 54 than 301 or 302 at Week 50: 1.90 vs. 0.88 or 1.08)

(302 W50 -0.105 SE=0.112 p=0.3482 95% CI=[-0.326 0.115];

301 W50 +0.073 SE=0.105 p=0.4837 95% CI=[-0.132 0.279]) despite the much larger sample size.

More of the 10 mg/kg difference from pooled placebo was in the APOE- subgroup in study 103 (-1.83 in APOE- vs. -0.77 in APOE+) which is the opposite of what was seen in study 302. There was no significant interaction between treatment and APOE though (interaction test for high dose vs. placebo at Week 54 by APOE, p=0.3746 but this would be underpowered). There was a very small difference in the APOE- subgroup in study 302, -.06 vs placebo on CDRSB at Week 78, and it was almost identical to the high dose effect in APOE- seen in study 301. Thus, pooled across 301 and 302 there is very little evidence of a high dose effect at Week 78 in the APOE- subgroup, although this group had 10 mg/kg dosing from the start of the study and less ARIA, so fewer dose interruptions.

Study 103 had a staggered multi dose design, so the 10mg/kg group is not completely concurrent with the full pooled placebo group and there could be a resulting bias. The analysis was acknowledged to be exploratory and the sponsor's result for CDRSB at Week 54 is sensitive to the handling of post randomization starting of concomitant AD medications in 103. The 10 mg/kg is not significant if data after starting post randomization AD medications are censored (p=0.09) or if a high dose outlier is excluded (p=0.10). This study doesn't seem very supportive due to sensitivity to handling of starting of AD medications, the staggered design without direct support of a dedicated randomization between 10 mg/kg and pooled placebo, interim analyses with some Biogen personnel unblinded, no effect in phase 3 at the Visit (Week 50) closest to the 103 study duration time (Week 54), and APOE- had a numerically bigger effect in study 103 than APOE+ which is the opposite of study 302 APOE subgroup results. The baseline adjusted Spearman correlation between Week 54 SUVR and Week 54 CDRSB change is also much larger in study 103 than the correlation at Week 78 in study 302 or 301 (0.60 [N=21] vs. 301 0.07 N=110 or 302: 0.13 [N=99]), which again calls study 103 into question. This illustrates that phase 1B and phase 2 studies have lower positive predictive value than phase 3 studies, i.e., they are less reliable.

A p-value < .05 doesn't necessarily reflect a clinically meaningful effect especially if there would be connotations of disease modification but the actual evidence supporting that is lacking and there is a failed study calling that p-value into question.

The final analysis of 302 is like a second interim analysis at 55% information since 45% of Week 78 data is missing in the final analysis. A 3 stage O'Brien Fleming boundary with interim analyses at 45, 55, and 100% would spend .010 by the 2nd stage. The critical value for stage 2 would be Z=2.68. The observed t statistic is 2.52. Therefore, there would be insufficient evidence to stop study 302 for early efficacy using the stopping boundary that the sponsor had prespecified in the protocol (i.e., hypothetically, if it had not already been stopped for futility).

Reproducibility of 302 is in question since 301 did not replicate it (e.g., "reproducible research" issue with p-values underestimating the probability of the null hypothesis).

5.2 Collective Evidence

The 2019 draft guidance on substantial evidence states that "poor execution can render a trial of any design to be not adequate or not well-controlled and, therefore, unable to provide substantial evidence of effectiveness. Examples of this include a randomized, double-blind, placebo-controlled trial in which unblinding is common due to an effect of the test drug, and where a modest treatment effect is found on a primary endpoint that is subject to bias when drug assignment is known (e.g., a physician global impression). In these cases, the trials might not be considered adequate and well-controlled." Both of these conditions are somewhat of a concern in this application although the outcomes have some objective items and there were separate rating and treating physicians used. Further the draft guidance states "Findings from other trials that are not consistent with the findings of the single positive trial would need to be considered collectively, and could weaken the overall strength of evidence."

Conditional power based on all ITT data with censoring after March 21, 2019 of meta analysis being positive if non-OTC patients by the time of futility were somehow able to complete is estimated to be .61. This kind of pooled analysis was the basis for the interim treatment estimate on which the conditional power calculations were based. Perhaps, this calculation does not adequately take into account potential heterogeneity between studies, but it summarizes all available phase 3 evidence and follows the spirit of the sponsor's original plan to use a pooled treatment effect estimate at the interim given the identical study design of studies 301 and 302.

5.3 Conclusions and Recommendations

The totality of the data does not seem to provide sufficient evidence to support the efficacy of the high dose. There is much inconsistency. There is only one positive study at best and a second study which directly conflicts the positive study. Both studies were not fully completed as they were terminated early for futility and had sporadic unblinding for dose management of ARIA cases which was much higher in the drug group. The Amyloid PET substudy data suggested a larger effect in APOE- which is the opposite of what was observed for the clinical outcome data in phase 3. Within the high dose group the correlation between Week 78 cerebellum SUVR change and Week 78 CDRSB change in study 301 or 302 is very small. Therefore, there is no convincing evidence of delaying clinical progression cognitive or functional: only a single positive timepoint (unreplicated and conflicted by a second study) and no delayed start design (termination for futility does not help with completeness or interpretability of long term follow up). The increased placebo progression post PV4 and smaller effect (on all 4 key endpoints) in APOE non-carriers who all got 10 mg/kg from the start of the study rather than like APOE+ having to wait until PV4, call into question the sponsor's assertion about intermediate dosing early (less 10 mg/kg doses) in the trial being a challenge. In addition, the low dose in study 301 was numerically better than the high dose despite having no 10 mg/kg doses and this comparison is supported by randomization. For these reasons, substantial evidence is lacking in this application.

This is a representation of an electronic record that was signed electronically. Following this are manifestations of any and all electronic signatures for this electronic record.

/s/

TRISTAN S MASSIE
05/10/2021 07:51:11 AM

KUN JIN
05/10/2021 08:14:58 AM
I concur with the review.

SUE JANE WANG
05/11/2021 11:13:20 AM

HSIEN MING J HUNG
05/11/2021 11:58:56 AM