

# **Appendix A**

## **A Critical Analysis of the AD2000 study**

**G. Silver, European Senior Medical Advisor, Eisai Ltd**

**R. Kay, Independent Statistical Consultant**

**R. Phillips, Health Economist, Goffin Consultancy Ltd**

## 1. Executive Summary

The AD2000 study has had a prejudicial influence on the present appraisal of Alzheimer's disease treatments by NICE. The Appraisal Committee, NICE secretariat and SHTAC place uncritical reliance upon this study but fail to recognise the very substantial limitations of the study, despite the widespread serious criticisms both in the form of published correspondence and as communicated to NICE by stakeholders.

This document lists the numerous failings of the study in terms of study objectives, methodology and conduct as well as in statistical interpretation and analysis. These may be summarised as follows:

- The investigators studied 2 different drugs (aspirin and donepezil) concurrently and any differential benefit is therefore difficult to determine.
- They recruited a mixed population of both AD patients and patients with vascular dementia (VaD). Not only is VaD outside the scope of the NICE review for AD drugs but also the results are therefore derived from 2 distinct patient populations, with differing characteristics.
- The authors set out to conduct a “large, simple real-life trial to produce reliable evidence”. They set out to recruit 3000 patients, yet only recruited 566 and yet continue to assert that the conclusions drawn remain valid, despite this much smaller population.
- Eligibility for the study was based on “uncertainty principle”. Those patients for whom there was a definite indication for (or against) donepezil were excluded from the study. This introduced a significant negative bias into the study with the recruitment of only those patients where there was a real doubt as to whether they would be able to benefit from donepezil.
- The design of this study inappropriately included the use of numerous washout periods of 4-6 weeks duration. This is highly unethical and shows little regard for the well being of the patients. Additionally, there was a high attrition rate that may have resulted from precipitous symptomatic decline due to the long washout period.
- There is no mention of monitoring or other quality assurance/control measures, staff training or compliance to study drug. It is therefore uncertain whether this study was performed to Good Clinical Practice standards (GCP) and whether variability between centres is likely to make the results uninterpretable.
- The statistical processes employed in this study would not have satisfied regulatory authorities and are unacceptable in a high quality study.
  - They used the method of randomised minimisation, which is strongly discouraged by the EU regulatory authorities (ICH E9, 1998).
  - They randomised subjects on 2 subsequent occasions, although the response seen before each randomisation may have adversely affected the response after re-randomisation. This is not acceptable in a crossover design.
  - Three interim analyses were performed and yet not accounted for in terms of the accompanying statistical penalty.

The authors conclude that donepezil patients were comparable to placebo patients in terms of the primary outcomes of this study, but fail to take into account the confidence

intervals associated with their results, that demonstrate, even on their data, a possible 6 month difference in delay to institutionalisation in favour of donepezil.

## **2. Introduction**

The preliminary conclusions of the Appraisal Committee with respect to Alzheimer's Disease patients as set out in the ACD, clearly place substantial reliance upon the AD2000 study. The study is referenced uncritically in five paragraphs of the ACD, in general to undermine the favourable results of other studies, even though AD2000 was not included in the meta-analysis for the cost effectiveness assessment. The reliance by the Appraisal Committee on AD2000, apparently disregarded the widespread and substantial criticisms of this study, both provided to NICE by stakeholders and published elsewhere. The failure to recognise any deficiencies in the study in presenting its results, appears to us to reflect a serious lack of balance in the way the evidence for this appraisal has been considered. In this context it is highly material that the Assessment Group has placed great weight on the importance of AD2000 and emphasise the significance of the findings of this study, while again failing to recognise the very serious limitations of the data. In the context of the approach followed by the Appraisal Committee in formulating the ACD, we are seriously concerned that the unbalanced views of the Assessment Group may have tainted the entire Appraisal process.

The study itself can be criticised on a number of levels but, in general terms, the design and conduct of the study as well as the subsequent interpretation of the results, suggest that it has no validity as a piece of scientific research in its own right. These issues in themselves provide sufficient grounds for exclusion from any proper review but even based on the inclusion criteria pre-specified by SHTAC, this study fails to meet the Assessment Group's own criteria for inclusion. Furthermore, quite apart from all of the deficiencies in the study design and methodology, the results themselves are unreliable as judged by the wide confidence intervals and no firm conclusions can reasonably be drawn from them.

The numerous failings of the study in terms of study objectives, methodology and conduct are set out below, together with the deficiencies in statistical interpretation and analysis. The limitations of the economic analyses are also detailed.

### **3. Study Objectives**

The publication of the study in the Lancet states that the objective of this study was to "answer various questions that were highlighted by NICE". The questions posed by NICE were stated in the original guidance document of 2001. It is therefore impossible that the study team set out with this objective in mind, since this study was designed and conducted from 1998 onwards. Indeed the study protocol states that the reasoning behind this study stems from uncertainty regarding clinical effectiveness in routine clinical practice, compounded by low levels of prescribing. Since this study was conceived less than 1 year after launch, it is hardly surprising that there was a limited evidence base following routine prescribing. Furthermore, restrictions in health authority funding, rather than "concerns about the moderate improvements in cognitive function tests" were largely responsible for limited prescriptions. This unfounded scepticism about the potential benefits of donepezil is therefore apparent from the outset and would certainly have contributed to negative bias in interpretation of the results and possibly even the design of the study.

3.1 Use of aspirin as well as donepezil means that the results are difficult to interpret.

An objective, as stated in the Lancet publication and trial protocol, was to determine if aspirin provided any benefit in the AD population. The trial protocol states that the objective was to “investigate whether aspirin really does delay disease progression”. Firstly, this was not a strictly defined AD population as discussed below, but the addition of aspirin resulted in 2 medicinal products being concurrently investigated in this heterogeneous population. This is highly unusual and raises questions as to how the differential results for the 2 compounds could be determined or how the effects of aspirin co-prescription or aspirin avoidance have influenced the study results, particularly as many of the participants had coexisting vascular dementia.

The results in respect of aspirin were not reported in the Lancet publication, although the authors indicate that they will be reported separately. So far as we are aware, the results in respect of the aspirin element have yet to be reported.

### 3.2 The study included patients with vascular dementia as well as Alzheimer’s Disease.

The inclusion of aspirin in the study was based on the planned recruitment of patients with Alzheimer’s disease with or without co-existing vascular dementia. The authors state that patients with any evidence of vascular dementia or concomitant disease had previously been screened out of donepezil trials and as such, they were keen to determine if “more clinically representative patients benefit from donepezil”. Patients with vascular dementia may certainly add to the patient mix presenting to clinical services but are a distinct population in terms of diagnosis, pathology and probable response to therapeutic and medical intervention. It is well known that vascular dementia patients behave differently from AD patients in terms of disease onset, progression and symptomatology, hence the existence of separate diagnostic criteria (NINDS-AIREN, ICD-10). In addition they have higher levels of co-morbidities and concomitant medication use (again indicated in the Lancet article where up to 53% had co morbidities, of which many were serious in nature including stroke and myocardial infarction as detailed in the Interim Report from Steering Committee, 2001). This associated cardiovascular and cerebrovascular disease will undoubtedly have a role in determining withdrawal rates from the study, independently of the response to donepezil and adds yet another variable to the interpretation of the study results. The pivotal studies for donepezil did indeed recruit a pure Alzheimer’s disease population and the licensing authorities granted an approval based on this population only. The efficacy for the compound in vascular dementia, or mixed disease, has yet to be demonstrated. The results of the AD2000 study therefore are derived from 2 different patient populations, each of which may influence the other positively or negatively, as well as from two different trial drugs. The ability to draw any reliable conclusions from this study is therefore limited or more likely, impossible.

The interim steering committee report of 2001 and the Lancet paper indicate that up to 18% of patients had vascular dementia and 4% had Parkinsonian symptoms. Dementia associated with Parkinson’s disease is yet another distinct diagnosis (DSM-IV) with characteristic features and again, a disease for which donepezil has not been proven to provide benefit. The objective of the AD2000 group was to conduct a large “simple study” to determine the potential benefits of donepezil in an AD population. This was not however an achievable task in a highly heterogeneous, mixed population with two investigational drugs.

In terms of the apparent divergence from the criteria identified by SHTAC for the inclusion

of clinical trial data, they defend their inclusion of this study by stating that only approximately 20% of subjects were diagnosed with vascular dementia. As stated above, no indication is given as to which, if any, criteria were used for VaD diagnosis in the AD2000 study. However the absence of a requirement for CT scans in the protocol suggests that the accuracy of diagnosis would be limited since neuroimaging evidence is central to all VaD diagnostic criteria. It is therefore highly probable that 20% substantially underestimates the proportion of VaD patients included in this study, for which reason alone, the study should have been entirely excluded from the assessment process.

#### **4. Study Methodology**

The following discussion outlines the major elements of the AD2000 study which refute the claim made by Gray et al. in a letter to the Lancet (October 2 2004) that AD2000 was a “methodologically rigorous study.” This is consistent with the overwhelming majority of comments submitted to NICE regarding this study and included in the Evaluation Report for this appraisal, point out the failings in study design that make any subsequent interpretation of results invalid.

##### **4.1 The study was too ambitious to produce reliable results**

The authors set out to conduct a “large, simple real-life trial to produce reliable evidence”. However, if the study design is far from simple and involves complex forms of randomisation. This level of complexity usually demands inclusion of a discrete population (rather than the mixed population of AD and VaD patients included in AD2000), appropriate assessments, and rigorous adherence to Good Clinical Practice (GCP) in order to produce reliable evidence. This study failed in these requirements (see paragraph 5.1 below) and yet continues to state its conclusions with an absolute level of certainty. The study ultimately carried out was neither simple nor large and for numerous reasons failed to produce reliable evidence: at the very least its results should be treated with a high degree of caution.

##### **4.2 Patients were only included if there was a real doubt as to whether they would benefit from donepezil therapy.**

Eligibility for the study was based on uncertainty. Those patients for whom there was a definite indication for (or against) donepezil were excluded from the study. The fact that patients who were likely to benefit were excluded means that this was not a representative population of Alzheimer’s disease patients but was instead biased against demonstrating benefit in patients eligible for treatment. It would appear that the AD2000 comparators were misguided in recruiting a heterogeneous population, for whom the evidence of benefit from donepezil was limited, and then attempting to demonstrate efficacy of the drug with inappropriate measures which were not intended for this patient population. It is not surprising therefore that this negative bias led to results in an under-powered study that were partially inconclusive and may not reasonably be generalised to mild to moderate AD population. It should be noted however that despite the planned recruitment of trial patients who were less likely to show benefit, the outcomes nevertheless showed significant global benefits in terms of cognition and function.

It is also relevant to note that the uncertainty principle for entry to the study was interpreted by different clinicians at different centres without any associated guidance to help with the

definition of inclusion/exclusion criteria. Indeed, the authors state that “even within one participating hospital different doctors may decide differently as to the categories of patient for whom the indication for donepezil is uncertain”. This can only add to the variability of recruited subjects and without very accurate and detailed recording of the reasons for inclusion, produce confounding results from patients who could be deemed entirely appropriate or inappropriate for inclusion by different doctors. Indeed the lack of inclusion and exclusion criteria may account for one of the concerns expressed by The Alzheimer’s Society. They comment in their response to the assessment report (October 2004) that the high rates of institutionalisation may indicate that “many people entered the study at the point of crisis”. This observation lends weight to the fact that the reasons for inclusion in the study are ill defined and the results are therefore difficult or impossible to interpret.

4.3 The primary endpoints are not those originally defined in the protocol.

The choice of endpoints in this study is also open to criticism.

Firstly, the endpoints have changed from those originally stated in the protocol, with no accompanying justification. The protocol states that the two primary endpoints were:

- Increase in disability as defined by either loss of two of four basic ADLs and/or loss of six of eleven instrumental ADLs
- The requirement for **formal domiciliary** or residential care

The Lancet article details that the second primary endpoint was in fact entry to institutional care but makes no mention of domiciliary care. The alteration in the second primary endpoint has therefore occurred post-hoc, no explanation for this amendment has been provided and no indication is given about this may have affected the power calculations or the results. This is not acceptable scientific practice and removes further credibility from the results.

The choice of endpoints is also concerning. The variability in access to institutional care, the differences in social service provision and the availability of informal care are all confounding factors which may independently affect the results of this study, making it difficult to judge a treatment effect.

Whilst it is accepted that functional improvements or stability are important in this progressive disease, no justification is given by the authors of AD2000 as to why the loss of 2 basic ADLs or 6 instrumental ADLs was appropriate in determining decline in this population. Lesser deterioration could still be considered clinically meaningful for example, the preservation of only 1 ADL such as the ability to eat.

Secondary endpoints included cognition, measured with the MMSE scale. The authors of AD2000 likewise set a high target for the MMSE assessment. Here they define a good response as an improvement of 4 or more points and define “no change” as including those patients who improved by an average of one point on the scale. These are inappropriately strict criteria, especially as the study cohort excluded most patients otherwise deemed suitable for donepezil treatment. The MMSE is known to be a fairly insensitive assessment but is used because of the high level of familiarity among clinicians. Therefore change demonstrable on this insensitive scale is likely to be significant and

there is certainly no rationale to suggest that an improvement of one point should be classified as “absence of effect”. No explanation or justification was provided by the study investigators for their definition of no change.

#### 4.4 The use of washout periods reduced the beneficial effects of donepezil therapy.

Benefits of donepezil therapy are typically lost after approximately 3 weeks without any treatment. Accordingly, the use of washout periods in the AD2000 study would have removed the benefits of previous treatment.

The AD2000 group were either unaware of, or disregarded, the wealth of Type II evidence available at the time of study design (Rogers SL, Farlow MR, Doody RS, Mohs R et al., *Neurology* 1998; 50:136-145, Doody RS, Geldmacher DS, Gordon B et al., *Arch Neurol* 2001; 58:427-433) to this effect. These data were available in the public domain at the time of design, and although the full manuscripts may have been published after the finalisation of study design, the strength of data that they highlight, ought to have been considered the subject of a protocol amendment. However the use of washout periods of between 4 and 6 weeks between each phase of the study meant that patients, on entry to each phase, had probably declined below the level of function attained at the end of the previous phase. Firstly this is highly unethical and shows little regard for the well-being of the patients and secondly the precipitous symptomatic decline may have led to excess withdrawals and hence bias in the study outcomes, as well a possible weakening of the blinding. It is not difficult for investigators, patients or their carers to detect rapid worsening over a short period and realise they have been on active treatment.

This flaw in the study has been the subject of numerous comments, including in the majority of submissions to NICE.

## 5. Study Ethics and Conduct

This study was considered by many to be both unethical in its design and conducted with little regard for Good Clinical Practice (GCP). The failure of the study team to give due thought and attention to these essential elements of study conduct causes significant concern and raises questions in respect of the reliability of the results. The following points provide further details of these issues.

### 5.1 The study was not conducted to standards of Good Clinical Practice

The study was not conducted in accordance with international standards of Good Clinical Practice (GCP). In fact, the protocol for AD2000 fails to make any statement about the standards to which the investigators adhered in running the study and these are not described in the Lancet paper. It is therefore unclear whether and, if so to what extent, any failures could have prejudiced the results of the study. This situation would be totally unacceptable if it were an industry-sponsored or registrational study and would not have passed an audit inspection for this reason. In these circumstances it is unclear why, this critical issue has not seemingly been recognised by NICE's appraisal committee. The following deficiencies in the protocol are of particular concern:

#### 5.5.1 Lack of monitoring.

There is no mention of monitoring or other quality assurance/control measures. It therefore must be assumed that no confirmation of the eligibility or existence of recruited patients took place. The investigators planned to recruit approximately 3000 patients from 50 centres and yet failed to comment on the methodology they proposed to employ to ensure adequate data quality. To recruit 60 patients per site and ensure adherence to protocol is a gargantuan task and one that needs considerable support and close follow up. In reality the study only recruited 565 patients, although some individual sites still recruited more than 70 patients and therefore the same concerns still apply. The Lancet article also cites the names of at least 120 investigators involved in recruiting these 565 patients. The variability that this would add to the data generated is enormous and has, again, not been adequately addressed in the reporting of the results.

The apparent lack of effective monitoring raises real questions as to the comparability of data generated at the various sites and whether they are meaningful at all.

#### 5.1.2 Extent of training not stated.

Despite the planned recruitment of 50 sites, no detail is given as to the training programme provided to ensure that the site staff were fully able to conduct the required assessments. In fact the protocol states that the selection of sites was based on merely expressions of interest rather than expertise. This is entirely unacceptable, particularly if there was an inadequate training of the less experienced sites and is likely to have resulted in inconsistent and incomplete assessment of patients, incomparable results and uninterpretable data.

#### 5.1.3 Compliance with treatment was poorly defined.

The Lancet article makes reference to the fact that compliance with study medication was assessed through the use of record cards and pill counts, but the protocol and the study manuscript fail to define the cut-off for determining non-compliance. It is therefore unclear the extent to which reported results are derived from a study population who actually were compliant with study medication.

#### 5.1.4 Use of concomitant medication was disregarded.

The protocol was keen to stress the importance of simple eligibility in this study, hence the limited inclusion and exclusion criteria. However the use of concomitant medication use could not, we suggest, properly be disregarded. Patients with significant co-morbidities were included in the study and the precise recording of the medications used to treat these conditions, is therefore essential to the accurate interpretation of results. No one would question the fact that significant deterioration in Parkinson's disease alone may account for decreased mobility and function and even entry to institutionalisation, independently of any dementia symptomatology. The associated change in PD medications would help reflect such deterioration. Likewise the use of neuroleptics with significant anti-cholinergic actions may dilute or even abolish any benefit seen with donepezil. It is therefore unclear why the AD2000 investigators did not consider these important factors and uncertain how differential use of concomitant medication

may have influenced the study results.

It is a widely held view that it was unethical to conduct this one-year placebo-controlled study at the point where donepezil had gone through the rigours of the regulatory approval process and was therefore deemed both safe and efficacious by global regulatory bodies. Had it not been for the restrictions imposed by many Health Authorities for accessing the drug at that time, undoubtedly even fewer patients would have entered. A quarter of patients had no exposure to donepezil for the entire year and would have certainly experienced significant deterioration over that time. The ethics of conducting placebo-controlled studies are often complex and less than clear-cut, but the decision to allow such a study to proceed must take into account the nature of the disease itself. In a progressive disease such as AD, decline in the double-blind phase is non-recoverable whereas in a stable disease, a delay to commencement of active treatment is generally less worrisome.

## 5.2 The study raises ethical concerns.

The study raises a number of ethical concerns including whether it was appropriate to carry out a placebo controlled study using a therapy licensed by regulatory authorities internationally as being safe and efficacious.

An additional issue with the use of placebo in this study design is the fact that some patients, who may have responded to active treatment, would then be switched to placebo at 3 months. This was unethical since the patients had no provision to be switched back to donepezil treatment. Any benefits achieved in the initial donepezil phase would have been washed out after 3 weeks, with the patient then continuing to decline over the course of the rest of the study.

While these ethical issues do not affect the overall results, we mention them here because they raise questions over the overall conduct of the study.

## 5.3 Other comments on study conduct

There appears to be little regard for the Data Protection Directive. The protocol includes a number of attachments, including the randomisation pad, where details are required of the patient's name and personal information. This is also evident in the adverse event forms and is indicative of widespread poor clinical practice.

No indication is given as to whether aspirin was regarded as an Investigational Medicinal Product (IMP) for the purposes of this study. The protocol and study publication seem to regard it as such in the description of study objectives and yet the protocol fails to describe the sponsor and investigator responsibilities that are associated with IMPs. This is particularly important in relation to adverse event reporting and the ongoing safety evaluation of IMPs that must take place in order to safeguard the well being of the study participants.

## 6. Statistical Considerations and Interpretation of Results

Commentators have raised significant concerns about the type of statistical methodology used in this study. It is clear that the statistical processes employed in this study would not have satisfied regulatory authorities with respect to determining a positive risk-benefit ratio. The problems outlined below throw further doubt on the validity of this study and provide further support for our view that it should not be included in this appraisal by NICE.

## 6.1 Method of Patient Allocation

The method of patient allocation is unclear. On p 2109 of the Lancet article there is reference to 'minimised randomisation'. Strict minimisation involves no element of randomisation; patients enter the trial and are placed, without a random element, into that treatment group which best corrects for existing imbalances in the variables (baseline characteristics) that are involved in the 'minimisation'. It is however possible to include a random element alongside this 'minimisation' where the patient allocation is not deterministic. In such cases there would be for example a 70%/30% randomisation to the relevant treatment group. The regulatory authorities use the term 'dynamic allocation' as a collective term for both minimisation and minimisation with the random element. In the ICH E9 document 'Statistical Principles for Clinical Trials' the regulatory authorities make it clear that they wish to see an element of randomisation. *ICH E9 (1998)*:

*'Deterministic dynamic allocation procedures should be avoided and an appropriate element of randomisation should be incorporated for each treatment allocation.'*

It is not clear if this element of randomisation was included in the AD2000 study. The AD2000 protocol criticizes this omission in previously reported data for donepezil and states that "...methods of randomisation and concealment are not reported...These are all potential sources of bias." The AD2000 group has failed to report accurately the methods of randomisation used in their study.

In addition, and subsequent to this ICH document, the regulatory authorities have become more concerned about dynamic allocation even when it includes a random element. The reason for their concern relates to the fact that there is some dispute about whether the properties of standard statistical procedures (for example, p-values) remain valid when such allocation methods are used:

*CPMP Points to Consider on 'Adjustment for Baseline Covariates' (2003):*

*'Dynamic allocation is strongly discouraged. However, if it is used, then it is imperative that all factors used in the allocation scheme be included as covariates in the analysis. Even with this requirement, it remains controversial whether the analysis adequately reflects the randomisation scheme. Applicants will be required to describe the sensitivity analyses they intend to perform to support the conclusions from the primary analysis. Without adequate and appropriate supporting/sensitivity analyses, an application is unlikely to be successful.'*

The types of supporting analyses to which the regulators make reference in this document are termed 'randomisation tests'. It is unclear from the AD2000 publication and protocol whether these were performed or whether the 'minimisation' variables have been included as covariates in the analysis, as recommended. This seems unlikely however since the only relevant statement indicates that "standard logrank methods were used to compare

rates of institutionalisation and progress of disability”.

In the context of any explanation for the apparent deficiencies in the randomisation used by the AD2000 investigators and for that reason alone, the study should be treated with substantial caution.

## 6.2 The Double Randomisation

One very concerning aspect of the design is that randomisation occurs twice. The authors claim that there are advantages. ‘First, patients who default from treatment in the first 12 weeks can be unbiasedly omitted from the long-term treatment comparison. Second, patients effectively contribute twice to the 12-week efficacy comparison.’

The first point unfortunately prevents the accurate evaluation of donepezil in a real-life, pragmatic setting since it leads to exclusion of those patients who withdraw early. This is directly contrary to the overall objectives stated by the AD2000 group at the outset of the study. The patient group randomised after the initial ‘12 week run-in’ are very much a selected group of patients who include some patients who are unable to tolerate donepezil. Subsequent extrapolation of the results of the long-term analysis, from this ill-defined group of patients randomised for the second time, to the wider population is consequently impossible. It would have been more straightforward and effective simply to randomise on a single occasion and use intention-to-treat methods of analysis to give a pragmatic answer to a pragmatic question.

Regarding the second point of double contribution to the efficacy comparison, the AD 2000 group have in effect taken the data from the first 12 weeks assessment, following the second randomisation, and put it together with the data from the ‘12-week run-in’. This essentially results in a two-period cross-over structure with 4 sequences:

placebo – placebo  
placebo – donepezil  
donepezil – placebo  
donepezil – donepezil

This does have the advantage of enabling within-patient comparisons in relation to the short-term endpoints. However there are numerous problems and assumptions. For example, there is likely to be a carry-over effect, where data in the second period (2<sup>nd</sup> block of 12 weeks) is potentially contaminated by the treatment given in the first period (1<sup>st</sup> block of 12 weeks, the run-in). No comment is given as to assumptions made by the AD2000 group in relation to this possible effect and it may well be that the modelling will have assumed that such a carry-over effect is absent. This would lead to inaccuracies since it has been well established in the pivotal studies for donepezil, that patients on active treatment respond above baseline levels within the first 12 weeks, whereas the placebo treated patients deteriorate, hence the observed treatment effect. Therefore entry to the second phase following the run-in period would include patients at very different baseline levels of function, dependent on the treatment received prior to this second randomisation.

A further problem in such trials concerns withdrawals in the first period which can bias the treatment comparison. In this study there were a substantial number of such drop-outs (36 subjects in the donepezil arm and 18 patients in the placebo arm).

### 6.3 The Statistical Methods of Analysis

This is another area of particular concern in this study. The methods of analysis are very unclear. Given the design, the most effective approach to the analysis of these data would have been to look at the data in two parts; the evaluation of the short-term endpoints from the 12-week cross-over section, and the evaluation of the long-term endpoints following on from the second randomisation.

All of the data however seems to have been combined in a single analysis. For example, logrank methods have been used to compare the primary (long-term) endpoints but these appear to have been applied to all of the data. This is evidenced by figures 2 and 3 of the Lancet publication which display 282 donepezil and 283 placebo patients at risk at time 0 - these numbers correspond to those randomised at the first stage to the 12 week run-in. This leads to the fundamental question of how the authors have accounted for the switch in treatment following 12 weeks and how has this been represented in the figures within the paper (Figures 2 and 3)?

In the same way the endpoints of MMSE, BADLS, GHQ-30, NPI and others have been analysed by including all of the data in a "multi-level" model. This model will contain a number of assumptions in order to be able to deal with such a complex design, to which the AD2000 group do not make reference even though they say they recognise the importance of adequate reporting of methods of analysis. They do however mention that the methods of Fleiss (ref 31) were also applied. This is very confusing because the methods detailed by Fleiss in this reference relate to meta-analysis and are unrelated to the situation in this study.

It is also apparent that at least 3 interim analyses were conducted during the course of the AD2000 (Bentham et al. Abstract presented at 8th ADRD, Stockholm, Sweden July 2002, Gray et al. Abstract presented at AD/PD 6<sup>th</sup> International Conference, Spain, May 2003). Reference is made to another interim analysis one of them in the Assessment report (Wessex DEC) prepared for the original NICE appraisal. It is well known that a statistical penalty is the consequence of interim analyses and no mention has been made in the Lancet study report as to whether this was taken into account.

### 6.4 Withdrawal Analysis

The AD2000 group have used logrank methods for evaluating time to institutionalisation. Whilst this is an acceptable approach it is important to know how they dealt with deaths and withdrawals. They fail to state whether they were treated as events or censored observations or potentially, as both.

Indeed the issue of withdrawals in this study is crucial and yet not addressed in the AD2000 report. The study design was such that treatment wash-outs, causing rapid symptomatic decline, and the extension from 12 weeks, at the point of consent, to a greater than 4 year study, after study initiation, would have been responsible for significant attrition. Furthermore the NICE guidance issued during the conduct of the study, would have resulted in substantial withdrawals since patients would have been keen to access donepezil treatment on the NHS without the associated risks and burden of wash-outs and study procedures. As such the remaining population in the study would be significantly biased.

This is reflected in the study outcome in which 48% of patients had discontinued after the 1<sup>st</sup> year and less than 20% remained by the end of the 2<sup>nd</sup> year. Indeed the population in the study fell from 565 patients at the outset to 51 in the 3<sup>rd</sup> year and to 4 patients in the 4<sup>th</sup> year. This vast attrition strongly suggests that numerous factors were influencing retention in this study. The authors make no attempt to categorise the type of study population remaining in the study, such that one could attest that the conclusions drawn are simple not applicable to any dementia population.

## 6.5 Sample Size and Extrapolated Conclusions

Much of the criticism that has been levelled at this study has focused on the remarkable shortfall in recruitment. For all of the reasons outlined above, patients, their carers and clinicians were unwilling to participate in the study. The AD2000 group were unable to recruit the number of patients they had calculated were required to produce meaningful results. As such they conducted some retrospective power and sample size calculations in attempt to justify drawing conclusions from a far smaller study. Such post hoc rationalisation is not scientifically credible. The position consistent with the original power calculation is simply that 565 patients were too few to produce any meaningful results.

There are additional issues that need to be considered when assessing the results from this small study. In a large trial it is far more valid to conclude that two treatments are similar on the basis of a non-significant p-value than it is in a small trial. This is because the power of a larger study is far higher. In a small trial there may in fact be large treatment differences but a non-significant result is obtained merely because the sample size is low. In this study the authors are essentially concluding that the two treatments, donepezil and placebo, are the 'same' with respect to 'entry to institutional care' and 'progression of disability' on the basis of non-significant p-values. Such a conclusion is simply not justified based on these data.

As ICH E9 points out:

*'Concluding equivalence or non-inferiority based on observing a non-significant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is inappropriate'.*

Again, as detailed in ICH E9, conclusions regarding equivalence should be based on confidence intervals and not on p-values.

In terms of the 'entry to institutional care' the p-value calculated is 0.4 with a relative risk of 0.97; the 95% confidence interval is (0.72, 1.30). Although the point estimate 0.97 is close to 1, the confidence interval has not discounted 0.72 (or indeed 1.30) as a potential value. A relative risk of 0.72 would suggest that the median time to institutionalisation in the donepezil group is approximately 0.72 times the median time in the placebo group. If the median time in the placebo group were 2 years then this would correspond to a median time of 1.44 years in the donepezil group, a difference of over 6 months.

Similarly for 'progression of disability' or 'entry to institutional care' the p-value is 0.7 with relative risk of 0.96 with a 95% confidence interval of (0.74, 1.24). The same arguments would apply with a potential reduction in the median time to event from 2 years in the placebo group to 1.48 years, again a difference of over 6 months.

It is therefore not correct to conclude similarity based on these data.

## 7. Economic Analysis

One of the primary objectives of this study was to determine delay to institutionalisation, although this was modified at some point during the course of the study as detailed above in the discussion concerning choice of endpoints. The economic analyses undertaken pose many, as yet, unanswered questions and raise points that require a greater degree of transparency and clarification. These are outlined below.

### 7.1 The study claims to consider a societal perspective.

The Lancet report (2004) states that the economic evaluation 'tested two hypotheses and adopted a societal perspective'. This evaluation did not seemingly consider a societal perspective as this would inherently take all costs into account, particularly productivity costs (time off work by patient or carer). The perspective taken by this economic analysis in fact appears to be that of the payer.

This is disappointing because in the context of Alzheimer's disease, we believe that the payer's perspective is particularly inappropriate because it disregards the additional very substantial burdens to society.

### 7.2 There is a discrepancy in the number of patients included in the economic analysis.

There appears to be a substantially increased number of patients reported in each group for the economic analysis compared to those in the clinical analysis over weeks 0-60 (**donepezil: placebo 997: 1013 compared to 283: 283 in the Lancet clinical report**). No explanation can be determined for this discrepancy, based on the information given in the Lancet article.

### 7.3 The heterogeneous population is likely to affect the primary endpoint of the study.

The recruitment of a heterogeneous population to this study may have a significant effect on the primary outcome of delay to institutionalisation. Detail has been given in the earlier sections about the very high likelihood of a mixed AD and VaD population in this study. Vascular dementia patients display a different symptomatology to AD patients with characteristic impairments in their ability to perform complex tasks along with the co-ordination of cognitive skills of planning, attention, concentration and self-control. All of these elements are central to the ability to perform activities of daily living, with the loss of functional ability being very closely correlated with entry to institutional care. It can be postulated therefore, that since the AD2000 population comprised a significant number of patients with vascular dementia, there may well have been an increased rate of patients requiring full time or institutional care. Furthermore VaD patients have higher rates of serious co-morbidity than observed in AD patients, in that they tend to have overt cardiovascular and cerebrovascular disease. These co-morbidities can independently increase the likelihood of entry to institutional care. No mention is made in the Lancet report as to whether these confounding factors were taken into consideration.

#### 7.4 The rate of progress to institutionalisation was higher than in other studies.

The report from the AD2000 Steering committee (February 2001) indicates that by week 60 of the study, a significant number of patients had already been institutionalised. This is somewhat contrary to other sources of data, that suggest that the delay to institutionalisation with acetylcholinesterase treatment can be in the order of approximately 2 years (Lopez et al., J Neurol Neurosurg Psychiatry 2001; 72:310-314, Geldmacher et al., J Am Geriatr Soc. 2003 Jul;51(7):937-44.). The Alzheimer's society also commented in their response to the Assessment Report that many patients entered this study on the "point of crisis". This body of evidence may suggest that patients had been delayed in their access to treatment so they could be entered into the AD2000 study, such that on entry, they were far more functionally impaired, hence their high rates of admission to care in the first year of the trial.

#### 7.5 Failure to report confidence intervals does not reflect that uncertainties in the data.

It is important to reemphasise the effect of the chosen methods of statistical analysis on the interpretation of the economic outcomes from this study. As indicated in the previous section, the reliance on the use of p values rather than confidence intervals may suggest absence of effect of donepezil treatment on delay to institutionalisation, when in fact, a difference between the active and placebo treatments does exist. Examination of the confidence intervals along with the point estimates indicates a possible delay to full-time care of 6 months. Impartial reporting of the results of this study should clearly state the associated uncertainties.

#### 7.6 Average costs were uncertain

Average costs were used based upon the study arm to which the patients were randomised. This raises two questions: firstly, it is unclear if there was any note taken of an individual patient's level of cognition or functional status as obviously the greater the severity of disease, the higher the use of resources and consequently, the cost of care. Secondly, while not stated, it can be assumed that there would be huge variation between patients in the level of resource consumption in any arm. No effort has been made to take this uncertainty into account.

As an aside it is interesting to note that the AD2000 group chose to determine cost-effectiveness in terms of "cost per day in a high level of disability". Despite the overall limitations of this study, the authors recognised the inherent problems associated with use of QALYs in this patient population. Unlike the SHTAC group, the AD2000 collaborative recognised that there was a "lack of any well-validated quality-of-life measure for Alzheimer's disease patients" thereby rendering cost/QALY measures meaningless.

### **8. Overall Conclusions**

The failings of the AD2000 study extend right from the fundamentally flawed and unrealistic objectives, to the inappropriate and perverse study design, the unethical and unacceptable method of conduct, the unsubstantiated and atypical statistical analyses and finally to the overtly unfounded and inaccurate conclusions that have been drawn.

In these circumstances, we are very concerned by the uncritical reliance placed by the Appraisal Committee upon these flawed data and the fact that this study may influence the treatment available to a highly vulnerable group of patients.