**Air pollution and dementia in older adults in the Ginkgo Evaluation of Memory Study: Online-Only Supplement**

**Table of Contents**

**1.** LexisNexis approach for reconstructing address history

Figure S1.Procedure for LexisNexis address extraction and merging of address histories.

**2.** Covariate ascertainment

**3.** Multiple imputation of *APOE* genotype and other covariates assessed at baseline

Table S1.Imputer variables incorporated into the multiple imputation procedure

**4.** Sensitivity analyses

i) Alternative air pollution exposures

Figure S2. HRs and 95% CIs describing associations between long-term PM2.5 exposure and incident all cause dementia, AD only, or mixed VaD/AD. Method (a) of extracting PM2.5 exposures as described in the manuscript is compared with the alternative averaging approaches (methods b & c). Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

Figure S3. HRs and 95% CIs describing associations between long-term NO2 exposure and incident all cause dementia, AD only, or mixed VaD/AD. Method (a) of extracting NO2 exposures as described in the manuscript is compared with the alternative averaging approaches (methods b & c). Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

ii) Adjustment for additional covariates

Figure S4. HRs and 95% CIs describing associations between long-term PM2.5 20-year average exposure and incident all cause dementia, AD only, or mixed VaD/AD. Method (a-20yr) is from the main manuscript, and the individual adjustment variables of bmi, alcohol, and physical activity were included separately. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

Figure S5. HRs and 95% CIs describing associations between long-term NO2 10-year average exposure and incident all cause dementia, AD only, or mixed VaD/AD. Method (a-10yr) is from the main manuscript, and the individual adjustment variables of bmi, alcohol, and physical activity were included separately. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

iii) Effect modification

Figure S6. HRs and 95% CIs describing associations between long-term 20-yr PM2.5 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 20-yr PM2.5 exposure and treatment assignment, sex, or study community. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

Figure S7. HRs and 95% CIs describing associations between long-term 10-yr NO2 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 10-yr NO2 exposure and treatment assignment, sex, or study community. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

Figure S8. HRs and 95% CIs describing associations between long-term 20-yr PM2.5 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 20-yr PM2.5 exposure and self-reported history of angina, heart attack, stroke, atherosclerotic cardiovascular disease (ASCVD), heart failure, high blood pressure, or diabetes. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

Figure S9. HRs and 95% CIs describing associations between long-term 10-yr NO2 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 10-yr NO2 exposure and self-reported history of angina, heart attack, stroke, ASCVD, heart failure, high blood pressure, or diabetes. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

**1. Extracting LexisNexis addresses and the algorithm for processing**

At the time of enrollment, GEMS participants lived near one of the four clinical sites (University of Pittsburgh/Pittsburgh, University of California-Davis/Sacramento, The Johns Hopkins University/Hagerstown, MD, Wake Forest University/Winston-Salem).1 Residential addresses from 2000 to 2008 were documented during the semi-annual follow-up visits occurring approximately every six months. If a GEMS participant moved during the study, we assumed the move date was at the midpoint between the originally recorded address start date and the previous follow-up visit date.

We submitted to LexisNexis2 a list of 3069 participant first and last names along with their 4050 GEMS addresses and address dates between 2000 and 2008. No other identifying information was submitted to LexisNexis, and we asked them to provide us with up to twenty matches for each participant. The query returned 10,966 additional LexisNexis addresses along with a “first seen” and “last seen” date. Many of these addresses were duplicates, some with spelling differences, and there were also many address date windows that overlapped within a participants’ LexisNexis history. Figure S1 diagrams the process of collecting and reconstructing address histories.

LexisNexis professional support recommended that we consider the following while determining the sequential address histories: a) give GEMS addresses priority over LexisNexis addresses, b) treat the “first seen” address date as more reliable than the “last seen” date, and c) undervalue or ignore addresses containing less than six months of history. After visually inspecting a random sample of address histories for two hundred participants, our team developed a set of specific rules that were used to construct likely sequential address histories. These rules, outlined in the steps below, were coded using the R statistical computing language3, which allowed us to establish a reproducible sequence of residential histories prior to the year 2000 for each GEMS participant. The following six-step algorithm was carried out separately for each GEMS participant.

* Step 1 - Fuzzy Matching of Addresses: To merge the dates of identical GEMS and LexisNexis addresses, we also had to identify addresses that were likely the same address, but contained mild character differences. While we worked to clean both GEMS and LexisNexis addresses, there were still some leftover inconsistencies in address numbering and spelling. We used the R function “adist” 3 to compute the generalized Levenshtein distance4 between two addresses. This is the distance between two words based on the number of character edits it would take to make one address equivalent to the other, either through insertion, deletion, or substitution. We tested many different hypothetical address errors and how those errors would affect the overall address history to be produced by the algorithm, and decided on a cutoff of five. That is, if two addresses could be made identical by changing at most five characters, then the two addresses were treated as being identical.
* Step 2 – Remove addresses with less than six months of residency: Any LexisNexis addresses with less than six months of occupancy were not used in the final histories.
* Step 3 – Remove PO Boxes: We eliminated PO Boxes, Rural Route, or Box addresses from the final address histories since we were focused on precise location-based exposure metrics. These zip-code level addresses were still utilized in the initial LexisNexis query, however, allowing us to produce historical residential locations for those individuals that would have otherwise been missing.
* Step 4 – GEMS addresses take priority over LexisNexis addresses: Because GEMS addresses were verified and updated by participants or a proxy throughout the study, the GEMS addresses were treated as being more reliable than the LexisNexis addresses. If there were inconsistencies between GEMS and LexisNexis during the study period 2000 to 2008, only GEMS addresses were used. Furthermore, a GEMS address matching an historical LexisNexis address was prioritized over an historical LexisNexis address that did not match GEMS.
* Step 5 – Only use first-seen LexisNexis dates: LexisNexis professionals advised that the “first seen” dates were more reliable than the “last seen” dates. This could be due to an individual neglecting to promptly change their address at the time of a residential move, for example, with the division of motor vehicles or their voter registration. We therefore only used the “first seen” dates to construct the final address histories (except we still used the “last seen” dates for eliminating addresses in Step 2).
* Step 6 – Addresses with the same first-seen LexisNexis dates are used simultaneously: In rare cases, two or more LexisNexis addresses that were not GEMS addresses had identical “first seen” dates. With no way of deciphering which of these two addresses may have been the true residential location, we utilized both addresses simultaneously when quantifying the exposure metrics.

The algorithm described above and coded within R ensured that the address history construction was clearly documented and replicable.



**Figure S1. Procedure for LexisNexis address extraction and merging of address histories.**

**2. Covariate ascertainment**

Neighborhood Deprivation Index: Using residential history information, we also characterized the neighborhood social environment for each participant at the U.S. Census tract level between 1980 and 1999. Specifically, through an existing principal components analysis starting with 21 variables, a neighborhood deprivation index (NDI) was constructed from a weighted linear combination of the most informative seven variables from the 2000 U.S. Census: percent with Bachelor degree; percent in managerial occupations; median home value; percent with at least a high school education; percent interest, dividend, or rental income; median household income; and percent with annual household income greater than $50,000.28 Larger values of NDI are associated with higher deprivation levels.

Occupational History: A health habit questionnaire was given to GEMS participants during the baseline visit. The question asked “What was your usual occupation?”, out of which they were to choose one: 1) Professional/technical/managerial/administrative, 2) Sales/clerical service, 3) Craftsman/mechanic, 4) Machine operator/laborer, 5) Farming/forestry, 6) Domestic work/childcare, 7) Homemaker, or 8) Other. Items 3-6 were combined into a new category titled “Manual Labor” due to small frequencies. The new occupation variable had the following final category labels: 1) Professional, 2) Sales or clerical, 3) Manual Labor, 4) Homemaker, or 5) Other.

Physical Activity: Participants were asked to report the type and frequency of physical activities they engaged in during the past year at visits 4, 6 and 8 (1, 2 and 3 years after the baseline visit). Questionnaires collected closest to the baseline visit were used. We created a physical activity measure based on the approach of the validated Physical Activity Scale for the Elderly (PASE)8 questionnaire designed to assess physical activity in adults age 65 years and older. The total physical activity score was calculated from the reported frequency of engaging in the following seven activities: gardening and yard work, walking, volunteering, assisting family or friends, hunting, fishing or camping, babysitting and shopping. Response options were “never or less than once a month”, “once a month”, “few (2-3) times a month”, “once a week”, “few (2-3) times a week” and “every day”. Responses were assigned a score of 0-5 and then summed to a total score ranging from 0 to 35 with higher scores indicating greater physical activity. Missing values were imputed by taking the average of the non-missing responses within each participant if at least three of the questions had complete data.

Cardiovascular Disease Indicators: A medical history questionnaire was given to GEMS participants during the baseline visit. Seven binary variables came from the yes/no question: “Has a doctor ever told you that you had”: 1) angina, 2) a heart attack, 3) a stroke, 4) heart failure, 5) high blood pressure, 6) diabetes. We created an additional binary variable as an indicator of atherosclerotic cardiovascular disease (ASCVD) for any GEMS participant who reported having angina, a heart attack, or a stroke.

**3. Multiple Imputation of *APOE* in GEMS**

Multiple Imputation by Chained Equations (MICE)5 was used to impute the missing *APOE* genotype data within R. After excluding the 23 participants with missing air pollutant exposure, the imputation model was built using all GEMS participants prior to all other exclusions to utilize all available information. We additionally and simultaneously imputed *body mass index* (n=9), *alcohol number* (n=38), *smoking* (n=46), *passive smoking percentage tertile* (n=60), *pack years* (n=189)*, Neighborhood Deprivation Index (NDI)* (n=2), *physical activity score* (n=84), *angina* (n=46), *heart attack* (n=35), *stroke* (n=37), ASCVD (n=21), *heart failure* (n=47), *high blood pressure* (n=63), and *diabetes* (n=34).

During the multiple imputation (MI) procedure, we generated *m*=10 complete data sets containing statistically plausible replacements for the missing data and combined those with the original observed data for use in the models.6 While a single imputation of the missing data provides unbiased parameter estimates, it can also result in an underestimate of the variability in the effect and artificially smaller P-values.6 Allowing the plausible replacement values to vary across many replicated data sets corrects for this underestimate. The variables used in the MI process are listed in Table S1, and we carried out separate MI procedures for the different averaging periods and air pollution metrics so that only one averaging period (5-year, 10-year, or 20-year) and one air pollution metric (PM2.5 or NO2) was included in each imputation procedure. The sequential adjustment models were fit for each of the m=10 complete data sets detailed. Final parameter estimates for each model were found by pooling the parameter estimates corresponding with each of the replicated m=10 data sets.7

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table S1. Imputer variables incorporated into the multiple imputation procedure**  **Variables included in every imputation procedure** | | | | | **Variables included depending on exposure of interest and averaging period** |
| Time to Dementia | Age at Randomization | Year of Randomization | Sex | Race | PM2.5 (5, 10, 15, or 20 year average) |
| Education | Treatment (Ginkgo/Placebo) | Clinic (Study community) | Number of Pack Years (quartiles) | Passive Smoking Percentage (quartiles) | NO2 (5, or 10 year average) |
| Smoking | Alcohol Number | All-Cause Dementia | Mixed/Vascular Dementia | Alzheimer’s Dementia | NDI (5, 10, 15, or 20 year average) |
| 3MSE Score | *APOE* genotype | Body Mass Index (bmi) | Alcohol number | Physical activity score |  |
| Angina | Heart Attack | Stroke | ASCVD | Heart Failure |
| High blood pressure | Diabetes |  | | | |

Abbreviations: 3MSE, Modified Mini-Mental State Examination; *APOE*, apolipoprotein E ε4 carrier status; NDI, neighborhood deprivation index; ASCVD, atherosclerotic cardiovascular disease

**4. Sensitivity analyses**

We conducted three sets of sensitivity analyses to assess i) effects of alternative averaging periods for the historical air pollution exposure metrics (figures S2 & S3), ii) effects of additional adjustment variables (figures S4 & S5), and iii) the interaction between air pollution metrics and site, treatment group, gender, diabetes, and cardiovascular disease indicators (figures S6, S7, S8, and S9). Missing covariates were imputed using the multiple imputation procedure described in section 3 of this supplement. Analyses & results are described in more detail below.

**i) Alternative air pollution exposures:**

Method “a” is the original approach for the results presented in the main manuscript. The 20-year PM2.5 and 10-year NO2 were taken from 1980-1999 and 1990-1999, respectively. This approach aims to quantify long term air pollution exposure among study participants, while allowing every GEMS participant to be compared over the same time period. NDI was also averaged over the same time period as the air pollution metric.

Method “b” incorporated air pollution averaging periods for PM2.5 and NO2 that were calculated twenty and ten years, respectively, prior to the year of randomization. For example, a participant randomized in 2001 would have a PM2.5 20-year average from 1981-2000 and an NO2 10-year average from 1991-2000. We also similarly evaluated 15, 10, 5, 4, 3, 2, and 1-year averages, prior to year of randomization, as well as the year of baseline examination. NDI was averaged over the same time periods as the respective air pollution metrics.

Method “c” was based on the maximum yearly air pollution exposure for each GEMS participant (1980-1999 for PM2.5 and 1990-1999 for NO2.). NDI was averaged over the same time period from 1980-1999 and 1990-1999, respectively.



**Figure S2. HRs and 95% CIs describing associations between long-term PM2.5 exposure and incident all cause dementia, AD only, or mixed VaD/AD. Method (a) of extracting PM2.5 exposures as described in the manuscript is compared with the alternative averaging approaches (methods b & c).** Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.



**Figure S3. HRs and 95% CIs describing associations between long-term NO2 exposure and incident all cause dementia, AD only, or mixed VaD/AD. Method (a) of extracting NO2 exposures as described in the manuscript is compared with the alternative averaging approaches (methods b & c).** Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

Figure S2 shows little difference between the two 20-yr averaging methods (a & b) of PM2.5, but HRs for all cause and mixed VaD/AD were attenuated for exposure periods of one to five or ten years relative to longer exposure periods. The maximum average air pollution exposure also had a significant effect on all cause dementia and mixed VaD/AD; however, the HRs were attenuated. For example, an IQR increase in maximum annual PM2.5 exposure was associated with a 14% higher risk of all cause dementia (95% CI: 2%, 28%) and a 23% higher risk of mixed VaD/AD (95% CI: 0%, 51%). In comparison, an IQR increase in mean PM2.5 exposure in the twenty years prior to enrollment in GEMS was associated with a 20% higher risk of all cause dementia (95% CI: 5%, 37%) and a 31% higher risk of mixed VaD/AD (95% CI: 2%, 68%). Since average air pollution declined consistently from 1980 to 1999, the maximum air pollution exposure could be considered a proxy for historical air pollution exposure.

We similarly see no discernable difference between methods a & b for the 10-yr NO2 averages displayed in figure S3. Shorter averaging periods of NO2 did not appear to alter the effects on all cause dementia, but shorter periods dampened the effect on mixed-vascular disease.

**ii) Adjustment for additional covariates:** These new adjusted models were based on the same average 20-year PM2.5 and 10-year NO2 described in the manuscript, from 1980-1999 and 1990-1999 respectively. NDI was again averaged over the same time period relative to the respective air pollution metric. The models included all of the covariates described in method a, but were also adjusted individually for body mass index (bmi), weekly alcoholic drink consumption, and physical activity. Figures S4 and S5 both display no meaningful differences in the observed relationships long-term air pollution exposure and all cause dementia, AD only, or mixed VaD/AD.



**Figure S4. HRs and 95% CIs describing associations between long-term PM2.5 20-year average exposure and incident all cause dementia, AD only, or mixed VaD/AD.** Method (a-20yr) is from the main manuscript, and the individual adjustment variables of bmi, alcohol, and physical activity were included separately. Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.



**Figure S5. HRs and 95% CIs describing associations between long-term NO2 10-year average exposure and incident all cause dementia, AD only, or mixed VaD/AD.** Method (a-10yr) is from the main manuscript, and the individual adjustment variables of bmi, alcohol, and physical activity were included separately.Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

**iii) Effect modification:** We observed no evidence of effect modification by treatment, sex, or study community (Figures S6 and S7). We similarly saw no evidence of effect modification by history of cardiovascular disease or diabetes (Figures S8 and S9).

****

**Figure S6. HRs and 95% CIs describing associations between long-term 20-yr PM2.5 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 20-yr PM2.5 exposure and treatment assignment, sex, or study community (NC, North Carolina; CA, California; MD, Maryland; PA, Pennsylvania).** Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

****

**Figure S7. HRs and 95% CIs describing associations between long-term 10-yr NO2 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 10-yr NO2 exposure and treatment assignment, sex, or study community.** Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

****

**Figure S8. HRs and 95% CIs describing associations between long-term 20-yr PM2.5 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 20-yr PM2.5 exposure and self-reported history of angina, heart attack, stroke, ASCVD, heart failure, high blood pressure, or diabetes.** Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.



**Figure S9. HRs and 95% CIs describing associations between long-term 10-yr NO2 exposure and incident all cause dementia, AD only, or mixed VaD/AD and the interaction between 10-yr NO2 exposure and self-reported history of angina, heart attack, stroke, ASCVD, heart failure, high blood pressure, or diabetes.** Age is the time scale in analyses adjusted for year of randomization, study community, treatment assignment, sex, education, occupational history, NDI, smoking status, pack-years smoking, SHS exposure, and *APOE* genotype.

**References**

1. DeKosky, S, Fitzpatrick A, Ives, D, *et al*. The Ginkgo Evaluation of Memory (GEM) Study: Design and Baseline Data of a Randomized Trial of Ginkgo Biloba Extract in Prevention of Dementia. *Contemp Clin Trials* 2006; **27**(3): 238–53.

2. LexisNexis. Available online: <http://www.lexisnexis.com/en-us/products/public-records.page> (accessed February 2019)

3. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2020; <https://www.R-project.org/>.

4. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 1966; **10**(8): 707-710.

5. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; **45**(3): 1-67.

6. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med*. 2013; **86**(3): 343-358.

7. Rubin, D. Multiple imputation for nonresponse in surveys. New York: Wiley & Sons, 1987: 75-78.

8. Washburn RA, Smith KW, Jette AM, Janney CA. The Physical Activity Scale for the Elderly (PASE): development and evaluation. *J Clin Epidemiol*. 1993;**46**(2):153-162.