The following ‘frequently asked questions and complaints’ refer to our pre-print entitled ‘The proportion of Alzheimer’s disease attributable to apolipoprotein E’, available at medRxiv: <https://www.medrxiv.org/content/10.1101/2023.11.16.23298475v3>

In it, we show that the common genetic variation underlying the ε3 and ε4 isoforms of the protein apolipoprotein E (apoE) accounts for the majority of Alzheimer’s disease (AD) and a large fraction of all-cause dementia. Some scientists have met these findings with incredulity, which we believe has arisen due to complex or unfamiliar concepts underlying the results not being well known or understood (even among plenty of epidemiologists and geneticists).

To help clarify aspects of the research, we have jotted down answers to some common questions and complaints that have been raised in scientific meetings, conversations or correspondence with peer-reviewers, journal editors, audience members, the media and in thorough interrogation by my wife (who, I hope it is fair to say, has been won over to our case).

This document is a work in progress, and I will expand as necessary. Don’t hesitate to be in touch if you have questions not covered here.

Dylan Williams, MRC Unit for Lifelong Health and Ageing at UCL, dylan.williams@ucl.ac.uk, March 2024

With input from Neil Davies and Emma Anderson, Division of Psychiatry, UCL

**QUESTIONS**

1. **What do your Population Attributable Fractions (PAFs) mean?**

It is helpful to think about these results in terms of absolute risk—i.e., how likely is it that an average person will develop AD in a usual lifetime (assuming life expectancy of 85 years)?

Of 100 individuals on average across the whole population (irrespective of *APOE* genotype), about 12 will develop AD by age 85.1

In contrast, of 100 ε2/ε2 carriers, let us suppose that there would be no more than 3 cases by this age.[[1]](#footnote-2) Thus, in a hypothetical world in which everyone had an ε2/ε2 genotype, 75% of all AD would not occur. In reality, if we could target some aspect of the molecular pathway from apoE to AD with interventions (e.g. therapeutics) to reduce the risk of ε3 and ε4 carriers to that of ε2/ε2 individuals we could reduce AD occurrence by 75%.[[2]](#footnote-3) Moreover, given the likelihood that some aspects of study design will have biased our findings downwards, we suspect the proportion may actually be higher than 75%, but will need to test this in other datasets.

1. **If the ε3 and ε4 alleles are causing most AD, why doesn't everyone carrying these develop the disease?**

Again, considering absolute risk, among ε3/ε3 individuals who comprise approximately 60% of the British population, somewhere between 7 – 12% will develop AD in their lifetimes.1 So many ε3/ε3 individuals may live their entire lives without developing AD. Even among ε4/ε4 individuals, who have a lifetime risk by age 85 of about 60%, the risk is not absolute – some individuals with very high risk live to very old ages without AD.

So how can the ε3 and ε4 alleles cause a large proportion of AD, and yet many individuals carrying them are not all destined to develop the disease?

To understand this, let's consider a helpful analogy where cause and effect are not debated: smoking and lung cancer. About nine in ten cases of lung cancer are attributable to smoking tobacco, i.e. this is a PAF of 90%, signifying that if we could remove tobacco smoking from the population entirely, 90% of lung cancer cases would not occur. However, among smokers, only about 15% will develop lung cancer in their lifetimes. So smoking causes most lung cancer, but most smokers don't develop the disease.

The reasons for this apparent paradox relate to individual risk when we are looking at common, complex diseases with many contributing causes, which include lung cancer and AD.[[3]](#footnote-4) For instance, lung cancer arises as a result of many contributing factors in addition to smoking exposure -- some of which will increase an individual's risk (like smoking and probably air pollution). In contrast, others will protect individuals, e.g. possibly genetic mutations that improve an individual's ability to repair damaged DNA within cells.

In epidemiology, we think about this concept in terms of a spectrum of risk for a disease (which we refer to as liability) and there being a threshold somewhere on this spectrum that if crossed, an individual will develop the disease in question. An individual will occupy a position on this scale as a combination of many factors (genetic and environmental) that affect risk throughout his or her life. If an individual's combined risk leaves them below the threshold for disease on this spectrum, the disease does not occur for them. If combined risk is above the threshold, bad news will be forthcoming.

In the case of smoking and lung cancer, people who smoke are all getting a very big nudge up the spectrum -- in other words, the whole range of total risk among all smokers shifts up the spectrum relative to the range of positions of non-smokers. The average smoker is much closer to the threshold than the average non-smoker, so it takes fewer other contributing factors to go against them before they cross it (Figure 1). However, despite the disadvantage that smokers experience with their positions on the spectrum versus the positions of non-smokers, there is a large range of positions that smokers can occupy on the risk scale due to differences in the multitude of other contributing causes. If the threshold for lung cancer is high, many won't reach this despite having heightened risk incurred by smoking, and they remain unscathed.[[4]](#footnote-5)

*Figure 1: depicting a risk scale (liability) for lung cancer among smokers and non-smokers*

A graph with red lines

Description automatically generated

*Note that i) the averages and ranges of positions among smokers and non-smokers differ enough that very few non-smokers have sufficient risk to cross the threshold for disease (although a small proportion still does), whereas a considerable number of smokers do; ii) most smokers don’t meet the threshold for disease despite their increased risk relative to non-smokers. So smoking is responsible for a large majority of people falling above the threshold, but most smokers remain disease-free. NB: units of the risk scale are arbitrary.*

In this way, it may be less accurate to call smoking the cause of 90% of lung cancer and more realistic to call it a component (or contributing) cause for lung cancer. There may be other ways by which we could reduce lung cancer risk by targeting other factors that would nudge people down the risk spectrum, e.g. reducing urban air pollution. If this alternate intervention applied to all smokers, their positions would shift down the risk spectrum by the extent to which the removal of the secondary cause limits their risk, and fewer of the smokers would then meet the threshold for lung cancer. So, there are more ways to reduce the risk of lung cancer among an average smoker than the obvious one (that is, persuading them to stop smoking). Nonetheless, the fact that there are other factors involved in lung cancer risk does not undermine smoking's position as the predominant cause. If we could persuade everyone never to start smoking (or to stop ASAP), we would be positioning all of these prospective smokers further down the risk spectrum, and this would prevent most disease from occurring because far fewer of them would reach the threshold. That is why it is extremely important to find contributing causes which lead to big shifts to the average positions where people lie on a risk spectrum.2

Returning to apoE and AD, the ε3 and ε4 alleles are moderate and strong contributing causes to AD risk, respectively. They are also very and moderately common, respectively. These facts mean that they place much of the population at much higher positions on the risk spectrum for AD than if they had inherited two copies of ε2. That is why exposure to variation in isoforms of apoE should be considered the predominant cause of AD, even though this will not cause everyone inheriting ε3 and ε4 isoforms to develop AD. There are undoubtedly other component causes of AD but without the large risk increases conferred by ε3 and ε4 that place individuals with the alleles much nearer the threshold for disease, most of these other causes would be largely or completely irrelevant for the majority of AD that occurs in the population.

1. **So if most people with ε3 or ε4 alleles still won't get AD, how would we know who to treat with apoE-related medications before some of them develop the disease?**

The last response has important implications for the understanding of preventive medicine. We cannot accurately predict whether individuals will develop the disease unless we are aware of, and can measure, a large extent of the contributing causes to a complex disease. In other words, returning to the concept of risk spectrums from the last answer, we would need to find out where a given individual sits on the spectrum. This is why knowing an individual's *APOE* genotype alone does not accurately predict who is at risk of AD.

In terms of public health, faced with a poor ability to target at-risk individuals precisely, an alternate strategy is to intervene more widely than necessary. A good example of this is the prescribing of statins to lower circulating LDL cholesterol and prevent the incidence of coronary artery disease (CAD). Many individuals who take statins would not develop heart problems with or without exposure to statins (the proportion which would not be above the disease threshold irrespective of the benefits of statins). But across a large number of statin users, lots who would otherwise sit above the threshold will have fallen below it owing to the drug's beneficial effects on circulating LDL cholesterol. In turn, a lot of heart disease is prevented across the population. Since we don’t know which of these groups people would belong to, we must give them somewhat indiscriminately to prevent heart disease among the population. Though like all drugs, statins have side-effects, so it helps if we can prescribe them based on known risk profiles and avoid over-prescribing to people that we are confident are low down the risk spectrum for CAD.

This notion that we must implement interventions among many individuals who may not benefit from them to prevent cases among the population, on the whole, has been termed the prevention paradox. It is helpful to keep in mind for guiding thinking about potential interventions. A good candidate for potential intervention is not too costly (because we will need to apply it widely) and should provide recognisable benefits to users with little to no detriment to them. For example, installing seat belts in cars is not expensive, prevents a lot of injuries and fatalities in road accidents, and causes little inconvenience to the wearers – even though for the vast majority of drivers who never experience life- or injury-threatening crashes, they are in effect entirely redundant.

This will be the challenge for translational medicine for apoE and AD. Given that the effects of ε3 and ε4 alleles are so pervasive within human populations, we will need to produce efficacious and cost-effective therapeutics that can be implemented widely and safely within the budgets of healthcare systems. They must be within the realms of acceptable tolerance regarding side effects when set against the considerable benefits they will confer to many users. This balance will vary according to *APOE* genotype, with ε4/ε4 individuals standing to gain greater risk reduction for a given level of the same intervention than an ε3/ε4 individual, for example.

1. **Is it realistic to expect us to prevent all the burden of disease that these results imply we can?**

Possibly not!

Without removing the ε3 and ε4 alleles from the population entirely, imagining a strategy to completely intervene at the right period(s) in the life course in the right cell type(s) to obviate all this excess risk is challenging.

However, note that we might still be able to make very meaningful gains in preventing or delaying AD in the population even if we cannot fully match the protection implied by the PAF for ε3 & ε4 carriage. If an intervention removed a fraction of the risk, it could still translate to large degrees of prevention. This is especially important when considering that short delays in the onset of a late-occurring disease like AD will lead to large falls in its prevalence. A shift in onset will allow many people to live their full lifetimes without the disease developing.

1. **How common are the ε3 and ε4 alleles?**

It became apparent to us during this research that the proportions carrying e3 and e4 alleles of *APOE* are often misunderstood and misreported by many dementia researchers and even by some geneticists. For instance, it is common among research articles to see a statement like "15% of the population have an ε4 allele". This mistake arises from a misunderstanding of allele frequency, a common statistic in population genetics. The allele frequency for the ε4 allele is indeed about 15% in individuals of European ancestry, but allele frequency is defined as the proportion of an allele among *all alleles* for that variant in a sample rather than a proportion of individuals in a sample that carry the allele. Since individuals inherit two copies of every allele -- one on a chromosome from each parent -- the proportion of people carrying an allele is substantially higher than allele frequency. For ε4, in the sample we used in our article, its allele frequency was 15.3%, with about 26% of individuals in the sample having inherited one copy and almost 3% having inherited two copies.[[5]](#footnote-6) This means the proportion of individuals exposed to at least one copy of ε4 (i.e. ε4 carriage) is just under 29%. Similarly, for ε3, its allele frequency in our sample was 76.8%, with about 59% of the sample having two copies and 95% of the sample carrying at least one.

This distinction between allele frequency and the proportions of individuals exposed to an allele may sound a little esoteric and irrelevant, but it is extremely important in calculating the burden of disease attributable to an allele. These calculations rely on the proportion of individuals exposed to a risk allele (and the additional proportion of individuals exposed to two copies of the same allele), so using allele frequency instead of the proportion of individuals exposed leads to underestimated PAFs, and some erroneous calculations in the past.

**COMPLAINTS**

1. **"You made the unusual choice of selecting ε2/ε2, an exceptionally rare genotype (0.6% of the current sample), as the reference standard. Typically, the most common genotype (ε3/ε3) is considered the reference, and ε4 (risk) and ε2 (protective) are examined in a dose-dependent manner.**

**The view of ε3 as harmful is unusual since it is the common genotype. Looking at a PAF of ε4 vs. an ε3/ε3 baseline would be perhaps more helpful and more comparable to existing literature."**

It is incoherent to regard ε3 as neutral for AD risk simply because it is the most prevalent allele. This misunderstanding has probably contributed to apoE’s role as a cause of the majority of AD being long underappreciated by the field and why our findings are important.

A PAF indicates the proportion of cases that would be avoided if a given risk factor was modified/removed from the population. A PAF for AD attributable only to the ε4 genotype using ε3/ε3 as the reference indicates the proportion of cases that would be eliminated if those with ε3/ε4 and ε4/ε4 had a similar risk to ε3/ε3 (and this calculation would be relevant only for a subset of the sample and its target population because ε2 carriage is not considered here). However, this is not the correct comparison because individuals with ε2 alleles have a lower risk than ε3 homozygotes. The correct question is what proportion of AD cases would be avoided if everyone had the lowest risk variants, i.e. using ε2/ε2 as the reference category. With a big enough sample containing many ε2 homozygotes (as we had in UK Biobank), there is no theoretical or statistical rationale for using ε3 as the base category, as doing so very substantially underestimates the contribution of variation in *APOE* to ADsince it implicitly ignores the contribution ε3 to AD risk.

Therefore, it does not make sense to model individuals of ε3/ε3 genotype as the reference group in risk calculations by default simply because it is the most common genotype, nor to quantify only the burden of AD attributable to ε4 relative to ε3 homozygotes.

As analogy, consider a scenario in which we are interested in the cardiovascular disease (CVD) burden attributable to excess adiposity. In middle-aged individuals in countries such as the US and UK, overweight (body mass index [BMI] 25-30kg/m2) is more common than both healthy weight (BMI 18-25) and obesity (BMI >25) – in UK Biobank, these proportions were 42%, 39% and 20% at baseline, respectively (figure 2).3 The suggestion in this question for our analysis would be the equivalent of stating that we should only quantify CVD that is attributable to obesity with reference to being overweight – since the latter is the most prevalent category in the population – whilst overlooking that overweight individuals in the population (the majority) have a higher risk of CVD than individuals of healthy weight.

*Figure 2: the most common category is not necessarily innocuous. Weight categories alongside APOE allele frequencies among UK adults aged 39-73 years, colour-coded by gradients of risk for cardiovascular disease and AD, respectively*

A chart with numbers and a bar

Description automatically generated with medium confidence

Another analogy would be the early investigations into the risk of lung cancer among smokers in the 1950s, when non-smoking among adults in countries such as the UK was rare (<5%) and studies were assessing disease risk principally among heavy versus moderate smokers – i.e. for exposure like this, the norm was to have heightened risk of lung cancer (relative to the rare few who did not smoke at all) and we would underestimate the true burden of disease attributable to smoking by not comparing risk among moderate and heavy smokers to non-smokers*.*

We appreciate that Individuals with an ε2/ε2 genotype have seldom been used as the reference group in dementia epidemiology, but as our analogies hopefully make clear, it is the correct modelling decision to quantify the full burden of AD attributable to the common molecular differences in apoE.

1. **“Your results don't make sense, AD is clearly a multifactorial disease”**

A very large PAF for the alleles in *APOE* does not contradict AD having a multifactorial aetiology. As described in answer 2 above, most ε3 carriers do not develop AD and even among ε4 homozygotes, penetrance may not be 100%. There appears to be many endogenous and exogenous factors that affect AD risk, either through interactions with apoE or independently. As per our response to question 2, the apparent contradiction of apoE accounting for most disease burden despite AD being affected by many other factors is related to complex diseases having various component causes, which combine to determine overall disease liability. However, not all component causes are equal in effect (differing in both strength and prevalence). A disease like AD is unusual among complex diseases in having one very strong and prevalent component cause (variation in apoE). If the excess risk experienced by ε3 and ε4 carriers was greatly attenuated or obviated entirely, then disease liability among these individuals would be so low that other component causes of AD would be mostly or entirely irrelevant – individuals experiencing them wouldn't meet the threshold for developing the disease. This is why variation in apoE is the predominant cause of AD.

1. **“Your findings are not accounting for confounding, reverse causation, effect modification (interactions) and mediators.”**

*Confounding and reverse causation*

The most important assumption regarding the calculation of PAFs is that they are based on estimates of how much exposure *causes* a difference in outcome risk. If we base our estimates on observational findings, this can be misleading because associations can arise from biases such as confounding (there are common causes of both exposure and disease) and reverse causation (the disease process alters the exposure).

However, confounding or reserve causation cannot affect our results. Confounding of germline genetic variants after transmission from parents to offspring at meiosis and conception is impossible. Furthermore, we know that variants in *APOE* will be inherited independently from other distal variants across the genome, i.e. Mendel’s law of independent assortment. This creates circumstances where genetic associations with outcomes are not confounded in a conventional sense (factors like smoking cannot determine an individual’s *APOE* genotype). Genetic associations can be prone to other potential biases, such as assortative mating. However, none have been shown to have an appreciable effect on the distribution of *APOE* alleles within populations such as the UK. For instance, there is no evidence for people choosing their partners based on their *APOE* genotypes.

The low probability of these biases impacting results, along with much experimental evidence linking variation in *APOE* specifically to AD (rather than nearby genetic variation), are why we are confident in referring to apoE as a causal factor for AD in our paper.

*Effect modification (interactions)*

The PAFs will incorporate effect modification and mediation by the pathway from *APOE* variants to AD or all-cause dementia risk. For instance, associations of *APOE* risk genotypes with AD may be higher in women than men and lower in individuals of black ethnicity than white, but our overall estimate from a sample of both men and women and the two ethnic groups averages these associations and tells us about the burden attributable to the genotypes in the population as a whole. This is a useful property when considering the overall burden of AD that could be prevented in the population if an exposure was removed, assuming that the characteristics of the sample are similar to that of the target population.

An important caveat on effect modification is that it may limit what we can infer about equivalent PAFs in other settings where the distributions of effect modifiers may be different, e.g. in other countries where ethnic groups differ.

*Mediation*

For example on mediation, let us reasonably assume that a large part of AD due to variation in apoE function is mediated by cerebral amyloidosis. Dissecting a PAF into the extent of AD caused by apoE via distinct pathways would not change the overall PAF estimate for specific *APOE* variants: mediation analyses would tell us what proportion of apoE’s effects acts via amyloid accumulation versus other pathways. All these mediating pathways would be, by definition, a result of the variants in *APOE* (and the differential effects of the apoE protein by extension). So, modelling the total effect of the exposure on an outcome irrespective of pathways provides a correct indication of the PAF of AD that is attributable to variants in *APOE*.

We note that identifying such modifiers and mediators of these pathways would be a worthy endeavour, but doing so is a whole research programme in its own right and goes beyond the headline findings we are publishing now.

1. **“You should not produce PAFs for risk factors that you cannot modify”**

 Another common complaint about PAF calculations is that they are futile when no intervention to modify the exposure in question exists. What is the point in showing the burden of disease that could be prevented if we have no means to intervene?

A more lenient suggestion is that there should be a hypothetical intervention to use with the exposure of interest to motivate this type of calculation. Some would say that studying PAFs for genetic risk factors cannot even be considered a hypothetical intervention because they make the common mistake of perceiving 'genetic risk' as unmodifiable. We rebutted this misperception with some brief commentary in our paper. There are strategies available to us that could, in theory (and hopefully practice), modify the risk of AD conferred by apoE -- either directly or by targeting other parts of molecular pathways that the protein is involved in.

Hence, our study's conclusion draws attention to an unmet priority: we should develop the means to modify this pathway, and researchers or funders have not given it proportionate attention (as far as we can tell—see next question).

1. “**Aren't we doing much on apoE - AD already?”**

While there has been much research to address the apoE pathway to AD, the extent of this appears to have been trivial over the last 30 years compared to competing hypotheses regarding AD aetiology – most notably, activity addressing the amyloid cascade hypothesis. For instance, as mentioned in our paper, in trials registered on clinicaltrials.gov, there is currently only one agent directly related to modification of *APOE* genotype / apoE function (LX1001), which represents *less than 1%* of all therapeutics being tested in humans for AD at present.4

We recognise (and indeed hope) that preclinical research addressing apoE and AD may have occurred in the pharmaceutical industry.[[6]](#footnote-7) Since much of this would occur behind closed doors, i.e. unpublished in academic literature, we would not necessarily be privy to the information. As far as we can see from public information, it appears that dementia researchers have not given the apoE pathway anywhere near sufficient attention, given that our findings and others show that there is the potential to make a huge step change in AD prevention and treatment by mitigating the effects of apoE. This follows periodic calls from other researchers over the years.[[7]](#footnote-8)

We also note that even where research into apoE's role in AD has or may have been taking place, the vast majority has focused on explaining how ε4 increases risk. Our findings show that it would also be of great value to understand how to combat the ill effects of ε3, which our analyses suggest contributes to about 30% of AD burden and possibly much more (since we suggest our results could be underestimating the PAFs for ε3 and ε4).

**FORTHCOMING …. QUESTIONS OR COMMENTS STILL TO ADDRESS**

**If 75% of disease occurs due to e3 and e4, what is responsible for the other 25%?**

**What is the relevance of other genetic risk loci, some of which also have large PAFs?**

**What does this mean for targeting amyloid (or tau) and anti-amyloid drugs?**

**What are the potential molecular mechanisms linking apoE to AD?**

If you have any ideas for others to add to this list, please email me at dylan.williams@ucl.ac.uk

**References**

1 Genin, E. *et al.* APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Molecular psychiatry* **16**, 903-907 (2011).

2 Rose, G. Sick individuals and sick populations. *International Journal of Epidemiology* **30**, 427-432 (2001). <https://doi.org:10.1093/ije/30.3.427>

3 Frank, P. *et al.* Overweight, obesity, and individual symptoms of depression: A multicohort study with replication in UK Biobank. *Brain, behavior, and immunity* **105**, 192-200 (2022).

4 Cummings, J. *et al.* Alzheimer's disease drug development pipeline: 2023. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **9**, e12385 (2023).

1. we suppose this because we haven't seen lifetime risk estimates for ε2/ε2 individuals specifically, probably because they are rare and hence hard to produce precise estimates for, e.g. in the paper cited above ε2/ε2 individuals are grouped with ε2/ε3 individuals. [↑](#footnote-ref-2)
2. A necessary extension of the assumption for a PAF is that we would need an intervention to modify risk conferred by apoE that does not alternately increase AD risk by other pathways, e.g. through detrimental side effects of drug use [↑](#footnote-ref-3)
3. some rare diseases occur solely due to one cause and the probability of developing them will be very high (even 100%) if exposed, e.g. Huntington's disease, which arises from inheritance of one genetic mutation. [↑](#footnote-ref-4)
4. Unscathed in terms of lung cancer, that is. Their smoking also puts them at greater risk of other conditions like heart diseases, which are much more common -- i.e. the threshold for disease is much lower, and proportionately fewer smokers escape this. [↑](#footnote-ref-5)
5. If a variant is in Hardy-Weinberg equilibrium (HWE), genotype frequencies (the proportions of a sample which are homozygous or heterozygous for an allele) can be calculated from allele frequency for alleles *p* and *q* using the HWE formula *p*2 + *2pq* + *q2* [↑](#footnote-ref-6)
6. to paraphrase a helpful comment on this by a colleague, there is a lot of money to be made if apoE-targeting drugs could be brought to market – a fact which will not have been lost on the pharmaceutical industry in the last several decades. [↑](#footnote-ref-7)
7. For example, <https://gladstone.org/news/gladstone-scientists-apoe-ideal-target-halting-progression-alzheimers-disease> [↑](#footnote-ref-8)