UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Egg consumption and mortality risk: application of target trial emulation and g-computation, and comparison with traditional longitudinal data analysis methods

Yohannes Adama Melaku

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Olivier THAS

2023
2024

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Egg consumption and mortality risk: application of target trial emulation and g-computation, and comparison with traditional longitudinal data analysis methods

Yohannes Adama Melaku

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Olivier THAS

Table of Contents

Abstract

Introduction: Egg consumption has been a subject of debate in nutritional science due to its potential impact on long-term health outcomes, particularly all-cause mortality. While some studies have suggested that high egg consumption may increase mortality risk, others have reported no significant association or even a protective effect.

Objective: This thesis aims to investigate the relationship between egg consumption and all-cause mortality over a 22-year follow-up period, using both the g-formula and Cox proportional hazards models to provide a comprehensive analysis.

Methods: A 22-year prospective cohort study was conducted involving participants from the China Health and Nutrition Survey (1993-2015). Dietary data were collected at baseline using 24-hour recall method, and participants were categorized based on their egg consumption: less than one egg per day and one or more eggs per day. The primary outcome was all-cause mortality. The g-formula was applied to estimate 22-year risks, while the Cox proportional hazards model was used to compare these findings. The analysis adjusted for potential confounders, including age, sex, urban residency, income, education, smoking, physical activity, body mass index, hypertension, and other dietary factors.

Results: The g-formula analysis revealed that individuals consuming less than one egg per day had an estimated 10% risk of all-cause mortality (95% CI: 9% - 11%), while those consuming one or more eggs per day had a slightly reduced risk of 9.1% (95% CI: 8.4% - 9.4%). The absolute risk difference was -1% (95% CI: -1.7% - -0.6%), indicating a statistically significant reduction in risk associated with higher egg consumption. The relative risk ratio was 0.90 (95% CI: 0.83 - 0.93), suggesting a 10% reduction in risk. The Cox proportional hazards model supported this trend with a hazard ratio of 0.91 (95% CI: 0.69 - 1.20), though with less precision.

Conclusion: This thesis provides evidence that higher egg consumption is associated with a modest reduction in long-term mortality risk. The g-formula's precise estimates, along with supporting evidence from the Cox model, suggest a potential protective effect of egg consumption on longevity when included as part of a balanced diet. These findings contribute to the ongoing discussion on dietary guidelines and the role of eggs in promoting health, highlighting the need for further research to confirm these conclusions.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Prof. Dr. Olivier Thas, for his invaluable guidance and feedback. I also extend my appreciation to Prof. Zumin Shi for providing the dataset used in this research. I would like to thank the participants of the China Health and Nutrition Survey for their contribution. Finally, I am grateful to my partner, family, and friends for their unwavering encouragement and motivation throughout my research. I would also like to extend my sincere thanks to Hasselt University for providing me with the opportunity to undertake my study. Their support has been instrumental in the completion of this research.

1. Introduction

Dietary risk factors are leading causes of morbidity and mortality (1). According to the Global Burden of Disease study, about 20% of all global causes of death are attributed to dietary factors. Diets low in fruits and vegetables, whole grains, and high in processed meats are the leading dietary risk factors causing high levels of morbidity and mortality (2). The economic cost associated with sub-optimal diets is high (3). Adherence to a healthy dietary pattern helps reduce intermediate metabolic risks of chronic diseases, such as inflammation and adipose tissue accumulation (4). For instance, consuming high levels of plant-based diet was associated with lower inflammation levels and body mass index (5).

In nutritional epidemiology, dietary exposure can be expressed in three forms: 1) dietary patterns; 2) food items and groups; and 3) nutrients (6). Dietary patterns consist of groups of food items determined by prespecified criteria (an a priori method), such as a healthy eating index (HEI) 2015, or through a data-driven approach (a posteriori), using multivariate analyses like principal component analysis which create the underlying latent dietary behaviour of people. Using dietary patterns considers, to some extent, the interaction of dietary components. This is particularly true in the a *posteriori* method (6). The second level of exposure involves food items and groups, based on individual items such as eggs, milk, and avocados, or combinations of similar taxonomy items like vegetables and fruits. Defining dietary exposure based on a single food item or group helps to clearly define the exposure (treatment strategy) (6). The third level of exposure is nutrients, such as vitamin D and omega-3 fatty acids. Although nutrients have a direct impact on body metabolism and health, assessing the quantity of dietary sources is challenging.

To define exposures precisely and design effective interventions, using specific dietary components, such as egg consumption, is more appropriate (7). Estimating the causal effect of latent exposure (in the case of dietary patterns) is challenging because the exposure is not well-defined (8). Additionally, designing interventions based on dietary patterns is impractical as they lack clear and precise exposure strategies. For instance, while it is possible to design interventions to increase the HEI score, people do not exclusively consume the food items that are components of the HEI (7). Furthermore, it is impossible for all people to have similar HEI scores. In summary, although comprehensive, using

dietary patterns as dietary exposure violates the assumption of consistency in estimating causal effects (9). Additionally, the stable unit treatment value assumption (SUTVA) might be violated (9). This is because the exposure levels (e.g., a person's score in the latent variables) are dependent on the overall distribution of the exposure and are ranked, which may affect the effect estimates. This means that potential outcomes for each person are likely to be affected by the treatment assignment of other individuals in the study. Although there are recent theoretical and statistical developments in estimating the causal effect of latent treatment in social sciences (10), its application in nutritional epidemiology has yet to be explored. Similar challenges arise when using dietary source nutrients as exposure. It is very difficult to precisely measure the amount of nutrients a person consumes in their diet (6). The nutrient content of the same food items can vary depending on how they were cultivated, processed, and consumed. It is less challenging to determine the effect of nutrients in supplements, where we know the exact amount, frequency, and duration of consumption, allowing us to clearly state the exposure specifications. Similarly, the effect of a nutrient on health outcomes can be estimated indirectly using the effect of a dietary component rich in that specific nutrient (6). Therefore, the best approach to estimate the causal effect of dietary exposure on health outcomes is to use dietary components (7).

There are several advantages to focusing on a single food item. It helps us define the precise intervention we are interested in, avoiding ill-defined interventions (9) from a causal perspective. In the context of randomized controlled trials (RCTs), we cannot design an intervention with dietary patterns as it violates consistency. Similarly, in the real world, designing interventions based on dietary patterns and nutrients (particularly those focused on dietary sources) is very difficult. Estimating the causal effect of dietary exposure is already a challenging task, and using dietary patterns and nutrients from dietary sources as exposures adds an extra layer of difficulty in framing the causal question, answering it, and interpreting the findings (7).

Even with a single food item, there are several challenges from a causal perspective, but these are less severe than those associated with using dietary patterns and nutrients. These challenges include the overall dimensionality of dietary exposure. Food intake is dynamic, varying in time, geography, and content (7). Focusing on specific food items ignores the interaction of dietary components and does

not account for overall dietary exposure. Furthermore, the dynamic nature of food consumption over time and geography is another challenge not specific to dietary components. Therefore, when framing a causal question to determine the effect of dietary exposure, it is important to consider the exposure specification to fulfill the assumption of consistency (7). This demonstrates the importance of explicitly applying the abovementioned assumptions and principles in nutritional epidemiology, such as taking egg consumption as a dietary exposure and determining its effect on all-cause mortality. In this thesis, measurement bias, which has been extensively investigated, is not the focus (11). Instead, the focuse will be on selection (collider) and confounding bias (9).

Eggs, widely consumed worldwide, have seen an increase in consumption from a median of 12.6 grams per day in 1990 to 21.0 grams in 2018 (12) (**Figure 1**). Eggs are sources of cholesterol, protein, choline, vitamin B7, vitamin A, and lutein and zeaxanthin (**Table 1**). Despite their rich nutrient content (13), high egg consumption has been associated with increased risk of intermediate health markers. For instance, in a meta-analysis of RCTs, higher consumption of eggs was associated with increased levels of low-density lipoprotein cholesterol (LDL-c) (14) but not with triglyceride (15).

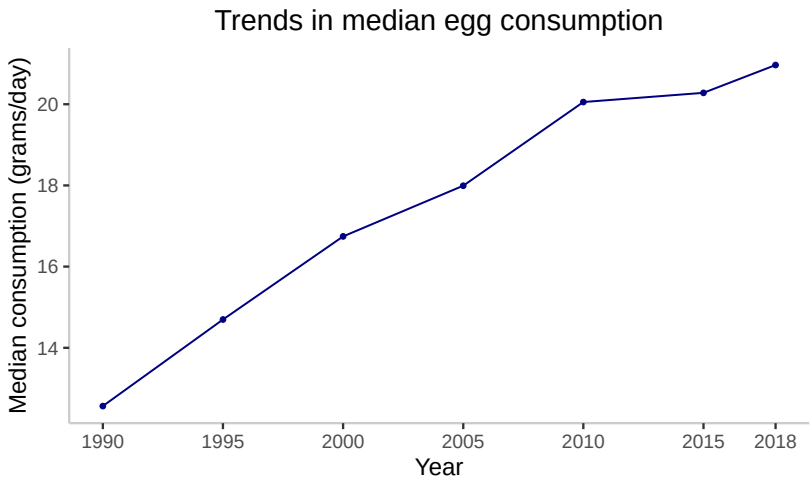

Figure 1. Global trend of egg consumption from 1990 to 2018 (12)

Table 1. Nutritional profile of egg (boiled one large egg)(13)

Nutrient contents	Unit	Quantity per 50 gram (one large egg)
Proximates		
Energy	Kcal	78
Water	g	37.31
Protein	g	6.29
Total lipid (fat)	g	5.3
Total carbohydrate	g	0.56

	Fiber	g	0
	Sugars	g	0.56
Minerals			
	Calcium	mg	25
	Iron	mg	0.59
	Magnesium	mg	5
	Phosphorus	mg	86
	Potassium	mg	63
	Sodium	mg	62
	Zinc	mg	0.53
Vitamins			
	Vitamin C, total ascorbic acid	mg	0
	Thiamin	mg	0.033
	Riboflavin	mg	0.257
	Niacin	mg	0.032
	Vitamin B6	mg	0.06
	Folate, DFE	µg	22
	Vitamin B12	µg	0.56
	Vitamin A, RAE	µg	74
	Vitamin A, IU	IU	260
	Vitamin E (α-tocopherol)	mg	0.52
	Vitamin D (D2 + D3)	µg	1.1
	Vitamin D	IU	44
	Vitamin K (phylloquinone)	µg	0.1
Lipids			
	SFAs	g	1.633
	MUFAs	g	2.038
	PUFAs	g	0.707
	Trans fatty acids	g	0
	Cholesterol	mg	186

Studies have also reported a link between egg consumption and long-term adverse health outcomes. A systematic review of 29 cohort studies with 2 million participants reported a relative risk of a 15% increased risk of heart failure associated with consuming one egg per day compared to no egg consumption. However, it found no effect on overall cardiovascular disease (CVD) outcomes (16). This association is consistent with the dose-response analysis conducted in the meta-analysis. In this review, the effect estimates of individual studies are inconsistent, ranging from protective to increased risk. Higher egg consumption is also associated with an increased risk of diabetes. The most recent meta-analysis, which included 16 prospective cohort studies, found a 7% increased risk of type 2 diabetes (RR=1.07; 95% CI: 0.99–1.15) for each additional egg consumed per day. However, individual studies reported highly heterogeneous results, with relative risks ranging from 0.62 to 1.45 (17).

It is not feasible to investigate the long-term effect of egg consumption with all-cause mortality risk using RCTs. Therefore, studies with observational designs remain the most important source of evidence. However, the research questions and methods used in these studies are highly inconsistent leading to inconsistent results. Most of the nutritional epidemiology studies do not use causal

approaches (7). In general, efforts to handle peculiar biases arising from the effects of confounding and selection bias in nutritional epidemiology are minimal, but this is less so for measurement bias (11). This problem starts with posing wrong causal questions accompanied by non-causal analyses and causal conclusions (7). For instance, most of the studies used only one time dietary intake data to define the exposure which is not optimal exposure specification. In addition, dietary behaviour specifically, lifestyle behaviour in general, are prevalent exposures, people are already have had the exposure for previously at the start of our study (7). On the other hand, the studies used different definitions of the exposure, i.e exposure specification is not consistent across the studies. Furthermore, the casual questions in those studies were not clearly stated. However, the interpretation of most of the findings from these studies were based on the assumption that causal methods were used.

Even though nutritional epidemiology strives to answer what nutritional interventions works for the community, causal inference theories, principles and methods are seldomly applied (7). To reach a relatively consistent conclusion on the effect of a dietary exposure on health outcomes from different studies in nutritional epidemiology, causal inference can be used as a tool. In this thesis, theories, principles and methods of causal inference were applied using an empirical data and compared with the most common used traditional statistical approaches. In this endeavour, the effect of egg consumption on all-cause mortality was determined using both the traditional and contemporary methods. Specifically, the aim of this thesis is:

1. To apply causal methods, namely target trial simulation (TTE) and g-formula, and determine the sustained effect of 22 years egg consumption on all-cause mortality
2. To compare the estimates with commonly used approaches in similar previous studie

2. Literature review and research gap

2.1 Egg consumption and all-cause mortality

The relationship between egg consumption and mortality outcomes has been the subject of numerous epidemiological studies, which vary in sample size, follow-up duration, and methodological approaches (**Table 2**). Across these studies, egg consumption was typically assessed through baseline food frequency questionnaires, with mortality outcomes primarily determined via death certificates or

national health registries (18-31). The reported risk ratios (RRs) for high versus low egg consumption demonstrated mixed results, ranging from 0.63 to 1.55, indicating both potential protective and harmful effects. These studies adjusted for a wide array of confounding factors, including age, sex, smoking, BMI, and various lifestyle and health-related variables. However, the inconsistency in findings across studies, coupled with potential biases related to dietary and outcome assessment methods, suggests that the causal relationship between egg consumption and mortality remains inconclusive. The diversity in study designs and populations highlights the complexity of isolating the effects of egg consumption from other dietary and lifestyle factors, necessitating further research with more consistent methodologies to clarify this association.

The evidence on the causal relationship between egg consumption and mortality is mixed, with studies showing both positive associations and no significant effects (**Table 2**). This variability underscores the importance of considering confounding factors, bias, and the heterogeneity of study designs when interpreting these findings (18-31). The diversity in adjustment factors across studies highlights the need for rigorous control of potential confounders to accurately understand the relationship between egg consumption and mortality. Additionally, variability in outcome assessment methods, such as the use of death certificates versus national registries, and differences in dietary assessment approaches introduce potential biases that can affect causal interpretations. The inconsistency in results across studies may be attributed to variations in study populations, follow-up durations, or unmeasured confounders, further complicating the ability to draw definitive causal conclusions.

Table 2. Characteristics of example prospective cohort studies that assessed the associations between egg consumption and all-cause mortality

First author, year (ref)	sample size, N	Follow-up period (years)	Dietary assessment	Outcome assessment	Egg consumption frequency or amount	Relative risk (95% ci) (high vs. low intake)	Adjustment factors
Kahn, 1984 (18)	27530	21	Baseline FFQ	Death certificates	<1 egg/wk Vs. 6-7 egg/wk	1.18 (1.00, 1.38)	Age, sex, history of heart disease, stroke, hypertension, diabetes, cancer, age at initial exposure to the Adventist Church, and smoking
Mann, 1997 (19)	10802	13.3	Baseline FFQ	Death certificates	<1 egg/wk Vs. >6 egg/wk	0.92 (0.68, 1.23)	Age, sex, social class, smoking
Nakamura, 2004 (20)	9263	14	Baseline FFQ	Death certificates	Seldom Vs. >2/d	Men. All-cause: 1.22 (0.68,2.99) women. All-cause: 1.53 (0.83, 2.80)	Age, serum creatinine, total cholesterol, blood glucose, BMI, systolic and diastolic blood pressures, use of blood pressure–lowering drugs, cigarette smoking, and alcohol intake
Qureshi, 2007 (21)	9734	10	Baseline FFQ	Death certificates	No or less than 1 egg Vs. > 6 egg/week	1.00 (0.90, 1.10)	Age, sex, race/ethnicity, systolic blood pressure, diabetes mellitus, serum cholesterol, cigarette smoking, BMI, and educational status
Djoussé, 2008 (22)	21327	20	Repeated FFQ	Death certificates	<1/wk Vs. ≥ 7/wk	1.23 (1.11, 1.36)	Age, BMI, smoking, history of hypertension, vitamin intake, alcohol consumption, vegetable consumption , breakfast cereal, physical activity, treatment arm, atrial fibrillation , hypercholesterolemia, and parental history of premature myocardial infarction
Sluik, 2014 (23)	258911	9.9	Baseline FFQ	Country-specific instruments	per 10 g	1.09 (1.06, 1.12)	Age, sex, prevalence of heart disease, cancer or stroke, educational attainment, diabetes medication use, and the following when there were no exposure variables: alcohol consumption, smoking, physical activity and underlying dietary patterns
Yu, 2014 (24)	61239	6.5	Baseline FFQ	National Registry and follow up home visit	Per 1 SD increase	0.95 (0.92, 0.98)	Age, education, income, smoking, alcohol consumption, multivitamin use, physical activity, BMI, waist-to- hip ratio, history of cardiovascular disease, diabetes, or hypertension, and total energy intake
Bongard, 2016 (23)	960	14.8	3-day food record	National Registry	Quartile 4 vs. 1	1.11 (0.70, 1.76)	Age, center, income, obesity alcohol, smoking, physical activity, chronic disease, diet score quality

Farvid, 2017 (25)	42403	11	Baseline FFQ	Family members, friends, or local health workers	0 Vs.26.4 g/day	0.88 (0.79, 0.97)	Age, ethnicity, education, marital status, residency, smoking, opium use, alcohol, BMI, systolic blood pressure, occupational physical activity, family history of cancer, wealth score, medication, energy intake
Guo, 2018 (26)	1781	22.8	Repeated FFQ	National health service central registry	0 ≤n≤1 Vs. n≥5 per week	1.08 (0.84, 1.38)	age , BMI, total energy intake , alcohol consumption , smoking , energy expenditure , social class, family history of myocardial infarction , diabetes mellitus, sugar intake , fruit consumption , red meat consumption, fiber
Virtanen, 2019 (27)	2641	22.3	4-d dietary records	National Registry	8 g/d Vs. 59 g/d	0.74 (0.63, 0.87)	age, examination year, and energy intake
Zhong,2019 (29)	29615	17.5	Baseline FFQ	International Classification of Diseases codes. and review of medical records, autopsy data, or both	0/d Vs. ≥2 per day	1.25 (1.04-1.52)	Age, sex, race/ethnicity, education, total energy, smoking, physical activity, alcohol consumption, use of hormone replacement therapy
Zupo,2020 (30)	2472	32	Baseline FFQ	National Registry	for one Frequency Per Day	0.63 (0.42 to 0.95)	Sex, age, BMI, education, smoking, comorbidity, wine, olive oil
Ruggiero, 2021 (31)	20562	8.3	Baseline FFQ	National Registry	>0≤ 1 Vs. > 4 egg/week	1.55 (1.17–2.06)	age, sex and energy intake

2.2 Use of causal inference methods in nutritional epidemiology

Associations between dietary exposures and disease outcomes are frequently observed, yet these associations do not inherently establish causation. Therefore, leveraging such findings to design public health and nutritional interventions can be inappropriate. Data analysis in nutritional epidemiology presents unique challenges and opportunities in implementing causal inference concepts and. As a core task of data science—alongside description and prediction—causal inference methods in epidemiology are increasingly employed to estimate causal effects of exposures on outcomes through observational studies (7). RCTs are typically considered the optimal method to establish causality due to the randomization of confounding variables. Estimating causal effects is pivotal in nutritional science as it delivers consistent evidence for actionable insights. While applying causal inference concepts and methods in nutritional epidemiology presents unique challenges (7), it remains a vital process if the evidence from nutritional epidemiological research is to be used to inform policy and decision-making. Most research questions in the field cannot solely rely on RCTs due to ethical considerations, feasibility constraints, or cost implications. As such, applying causal methods to observational data becomes imperative (32). In this thesis, target trial framework with g-formula were applied to estimate the causal effect of egg consumption on all-cause mortality using observational study from China.

2.2.1 Target trial emulation

In epidemiologic research, observational studies often serve as a practical alternative to RCTs, which are considered the gold standard for establishing causal relationships. However, observational studies are prone to biases due to confounding and other limitations. To bridge the gap between observational studies and RCTs, the concept of "target trial emulation" has emerged as a powerful methodological approach (33). This approach involves designing and analysing observational data in a manner that closely mimics the design and execution of an RCT, thereby aiming to draw more robust causal inferences.

Target trial emulation is an approach where researchers conceptualize the ideal RCT (the "target trial") that they would have conducted to answer a specific research question if ethical, logistical, or financial

constraints were not an issue (33). They then use existing observational data to emulate this trial as closely as possible. This involves clearly defining the trial's eligibility criteria, interventions, follow-up periods, and outcomes, and then applying these criteria to the observational data. The key components of target trial emulation are (33):

1. **Eligibility Criteria**: The first step in emulating a target trial is defining the population that would have been eligible for the RCT. In observational data, this translates to selecting participants who meet these criteria, ensuring that the study population is as comparable as possible to the population that would have been enrolled in the trial.
2. **Interventions**: In a target trial, participants would be randomly assigned to different interventions (e.g., treatment vs. control). In observational studies, however, treatments are not randomly assigned. Target trial emulation involves carefully identifying and classifying the exposures in the observational data that correspond to the interventions of the target trial.
3. **Follow-up and outcomes**: The follow-up period in the emulated trial must mirror that of the target trial. Researchers must define the start and end points of the follow-up and ensure that the outcomes of interest are measured in a way that aligns with the target trial's objectives.
4. **Confounding**: Since randomization is not possible in observational studies, target trial emulation requires rigorous methods to address confounding. Techniques such as propensity score matching, inverse probability weighting, or instrumental variable analysis may be employed to control for confounding factors and approximate the randomization process of an RCT.
5. **Analysis**: The analysis of the emulated target trial should follow the same principles as an RCT analysis. This includes using intention-to-treat or per-protocol approaches, depending on the research question and the availability of data.

Target trial emulation offers several advantages over traditional observational study designs. By aligning the study design with the principles of RCTs, this approach reduces biases and improves the validity of causal inferences. It allows researchers to leverage large-scale observational datasets to answer questions that would be difficult or impossible to address through RCTs alone. Additionally,

it provides a framework for transparency and reproducibility, as the emulated trial protocol can be explicitly described and evaluated (33, 34).

Despite its strengths, target trial emulation is not without challenges. The approach relies heavily on the availability and quality of observational data, which may not always capture all relevant variables or may suffer from measurement errors. There is also the risk of residual confounding, as not all confounders may be adequately accounted for. Furthermore, the success of target trial emulation depends on the researcher's ability to accurately define and operationalize the components of the target trial (33, 34).

Target trial emulation represents a significant advancement in the field of epidemiology, offering a rigorous and systematic approach to drawing causal inferences from observational data. By carefully designing and analysing studies to mimic RCTs, researchers can overcome some of the inherent limitations of observational studies and produce more reliable and actionable findings. As the availability of large observational datasets continues to grow, target trial emulation will likely become an increasingly important tool in epidemiologic research, providing valuable insights into public health interventions and outcomes.

2.2.2 G-methods

Intermediate confounding occurs when a confounder is influenced by prior exposure, creating challenges for traditional confounder adjustment methods, which typically hold confounders constant. These conventional approaches are inadequate for dealing with intermediate confounding for two main reasons: first, adjusting for an intermediate confounder can block part of the effect of the prior exposure; second, it can introduce collider bias by opening additional back-door paths between the exposure and the outcome. To address intermediate confounding, g-methods have been developed, which consider the observed distribution of intermediate confounders over time and across the population, rather than keeping them fixed. This allows for the estimation of marginal rather than conditional effects (9).

There are three main g-methods: g-estimation, marginal structural models and g-computation. g-estimation, using structural nested mean models, predicts the counterfactual outcome at each time point under the assumption of no further exposure, taking into account prior exposure and confounders (9). They offer reliable estimates of contrasts (such as differences and ratios) of average potential outcomes, under a broader and less restrictive set of identification conditions compared to traditional regression methods like linear, logistic, and Cox regression. Standard regression methods require the absence of feedback loops between time-varying treatments and time-varying confounders, whereas g-methods can accommodate such feedback. Marginal structural models aim to make the exposed and unexposed groups comparable by applying weights (often inverse probability of treatment weighting) to each observation, thereby balancing the distribution of confounders across groups. This allows for the calculation of the average treatment effect (ATE) through simple comparisons or unadjusted regression models. g-computation (or the parametric g-formula) uses a statistical model, such as a regression model, to predict potential outcomes for each individual with and without exposure, facilitating the calculation of treatment effects, though it depends on the correct specification of the statistical model (9).

2.2.3 G-formula

The g-formula models the joint distribution of the observed data to simulate potential outcomes across various exposure scenarios. Let's us assume the following causal diagram with two time points (**Figure 2**)

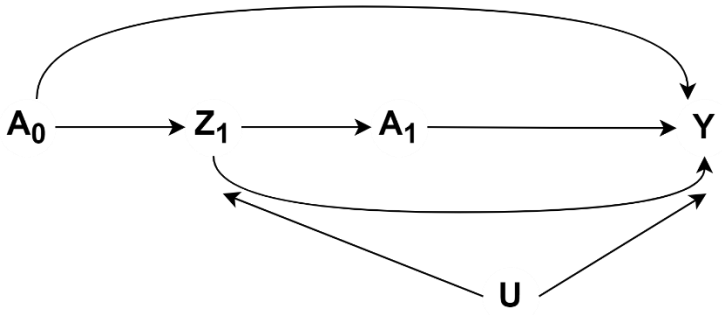

Figure 2. Direct acyclic graph depicting the association between treatment at time 0 (A_0), a confounder (Z_1) prior to the second treatment, treatment at time 1 (A_1), an outcome (Y) at the end of follow-up, and a common cause (U) [unmeasured] of the confounder and outcome.

Here, the average causal effect of always adhering treatment ($a_0 = 1, a_1 = 1$) compared to never adhering treatment ($a_0 = 1, a_1 = 1$) and average causal effect consists of the joint effect of A_0 and A_1on A is the focus. It is a marginal effect as it be an average of (or marginalise) over all participants in the population.

$$\varphi = E(Y^{a0=1,a1=1}) - E(Y^{a0=0,a1=0})$$
$$= E(Y^{a0=1,a1=1} - Y^{a0=0,a1=0})$$

where $E(.)$ is taken as the target population from which sample is drawn. $Y^{a0,a1}$ shows a potential outcome value had the exposures been set to a_0 and a_1. This potential outcome is different from the observed outcome.

Alternatively, we might want to estimate this effect based on specific values of another covariate. A conditional effect would occur, for instance, if one were particularly interested in how the effect varies with the covariate Z_1. When accurately modelled, this conditional effect provides an answer to: what is the effect of $A_0$and A_1 among individuals with Z_1=1 compared to those with Z_1=0?

We can factor the joint density by conditioning each variable. For instance, if $f(.)$ represents the probability density function, conditional probabilities is (9):

$$f(y, a_1, az_1, a_0) = f(y|a_1, az_1, a_0)P(A_1 = a_1|Z_1 = z_1, A_0 = a_0)$$
$$P(Z_1 = z_1|A_0 = a_0)P(A_0 = a_0)$$

We are interested in the marginal mean of Y that would be observed if A_0 and A_1 were assigned specific values a_0 and a_1, respectively. We first express the conditional mean of Y in terms of its conditional density.

$$E(Y) = \sum_{a1,z1,a0} E(Y|A_1 = a_1, Z_1 = z_1, A_0 = a_0)\, P(A_1 = a_1|Z_1 = z_1, A_0 = a_0)$$
$$P(Z_1 = z_1|A_0 = a_0)P(A_0 = a_0)$$

The average of potential outcomes is:

$$E(Y^{a0,a1}) = \sum_{a1,z1,a0} E(Y|A_1 = a_1, Z_1 = z_1, A_0 = a_0)$$
$$P(Z_1 = z_1|\, A_0 = a_0)$$

The above equation is the g formula (9). This method of calculating the g-formula is known as nonparametric maximum likelihood estimation (9). Parametric regression models are used in population health research that involve numerous covariates.

Assumptions: The average causal effect is defined as the difference between two averages that would be observed if everyone in the population were exposed (or unexposed). However, it is not possible to directly measure the averages because some individuals will be exposed while others will not. Therefore, part of our task is to justify using averages from subsets of the population as representatives of what would be observed across the entire population. This is achieved by making three key assumptions.

C***onsistency*** (9) enables us to equate the outcomes of those who received a specific exposure value with the potential outcomes that would occur if everyone were exposed to that same value.

$$E(Y|A_1 = a_1, A_0 = a_0) = E(Y^{a0,a1}|A_0 = a_0, A_1 = a_1)$$

The validity of this assumption does not change based on the analytic method used. Instead, its validity is determined by the characteristics of the exposure strategy. With consistency, we are able to identify average causal effect.

The second assumption is **exchangeability** (9). Exchangeability means the potential outcomes under exposures $a_0$and a_1 ($Y^{a0,a1}$) are unrelated of the observed exposures A_0 and A_1. Exchangeability assumption can be applied within specific levels of past covariate values

$$E(Y^{a0,a1}|A_1,\ Z_1\ ,\ A_0) = E(Y^{a0,a1}|Z_1, A_0)$$

$$E(Y^{a0,a1}|A_0\ ) = E(Y^{a0,a1})$$

Not conditioning for Z_1 will lead to uncontrolled confounding of the effect of A_1, creating a dependence between the A_1 and the potential outcome. But, applying standard methods to adjust for Z_1 (such as conditioning in a linear regression model) would partially block the effect of A_1 through

Z_1 and could introduce selection (collider) bias on the effect of A_1 through U. This challenge is what g formal was designed to overcome (9).

The third assumption is **positivity** (9). Positivity is satisfied when there are both exposed and unexposed individuals across all levels of confounders and prior exposures, which can be assessed empirically. This condition requires that $0 < P(A_1 = 1|Z_1 = z_1, A_0 = a_0) < 1$ and $0 < P(A_0 = 1) < 1$. Additionally, this assumption must be valid for all values of a_0 and Z_1 where $P(A_0 = a_0, Z_1 = z_1) > 0$. This ensures that consequences are not defined in levels of a0 and z1 that do not appear.

Given these three assumptions, hypothetical observational study can be viewed as similar to a serially randomized trial, where the exposure is initially randomized at baseline and then re-randomized at time 1, with the probability of exposure depending on Z_1. Under these conditions, g-formula can be applied to observational data to estimate counterfactual outcomes.

2.2.4 The causal question of this thesis

The causal question that this thesis will answer is: "*what is the effect of long-term consumption of one egg or more per day on the 21-year risk of all-cause mortality?*" The estimate from the causal analysis was compared with naïve (most commonly used method) estimates of the effect of egg consumption on all-cause mortality.

3. Methods

3.1 Description of the data

3.1.1 Study design and population

Longitudinal data was used from the China Health and Nutrition Survey (CHNS), an open prospective cohort study representing nine provinces in China (35). Data collection was conducted over eight waves, spaced two to three years apart, from 1991 to 2011. Households were chosen using a multistage random-cluster sampling method, and all household members were eligible to join the study. Throughout the study period, from 1991 to 2011, a total of 44,453 participants were involved in at least

one wave (36). After excluding those who were ineligible, the baseline sample consisted of 10,154 individuals. The response rates for participants who continued from previous waves into subsequent surveys were approximately 88% (37). The number of participants by year: 10154 (1991); 9932 (1993); 11,141 (1997); 12,557 (2000); 12,502 (2004); 15,1666 (2006); 15,670 (2009); and 19,041 (2011). The CHNS received approval from the institutional review boards of both the University of North Carolina (Chapel Hill, NC, USA) and the National Institute of Nutrition and Food Safety (Beijing, China). Informed consent was obtained from all participants prior to their participation in the survey.

3.1.2 Dietary data

Detailed descriptions of dietary measurements can be found in another source (38). In summary, dietary intake data were collected for each individual using a 24-hour dietary recall method over three consecutive days during each wave. Interviewers measured and recorded all food available and any food waste at home at the beginning and end of the three-day period (35-38). This information was combined with the dietary recall data to evaluate individuals' dietary intake. The Chinese Food Composition Table was used to analyse daily food consumption (in grams) and calculate nutrient intake. For detailed analysis, foods and nutrients were organized into 34 and 21 categories, respectively. This study specifically focuses on one dietary component: eggs.

3.1.3 Mortality

The primary outcome of the study was all-cause mortality. To track this, face-to-face surveys were conducted with participants and their family members during each follow-up wave. Family members provided the exact date of death during these follow-up surveys. The follow-up time was determined based on the date of death or the last survey date, whichever occurred first (35, 36).

3.1.4 Covariates and confounding variables

The models included the following baseline covariates: **socio-economic variables**–age, sex, urban, income, residency, education; behavioural: smoking, alcohol drinking, physical activity; **metabolic variables**–hypertension, body mass index; **disease conditions**–cardiovascular disease, and diabetes; **dietary components**–vegetable, fruit, whole grain, legumes, nuts, meat, fish, milk, and beverages; **energy intake**–total energy intake. The models also included the following time-varying covariates:

income, education, smoking, physical activity, alcohol intake, BMI, hypertension, diabetes, cardiovascular disease, vegetable, fruit, whole grain, legumes, nuts, meat, fish, and soft drinks and toral energy intake. All models included indicator variables for period of follow-up, baseline covariates, and time-varying covariates measured in the previous questionnaire. See **Supplementary Table 1** for details on the variables and models used. **Figure 3** provides a causal directed acyclic graph (DAG) depicting the assumed relationships between egg consumption and all-causa mortality in two follow up periods only for readability.

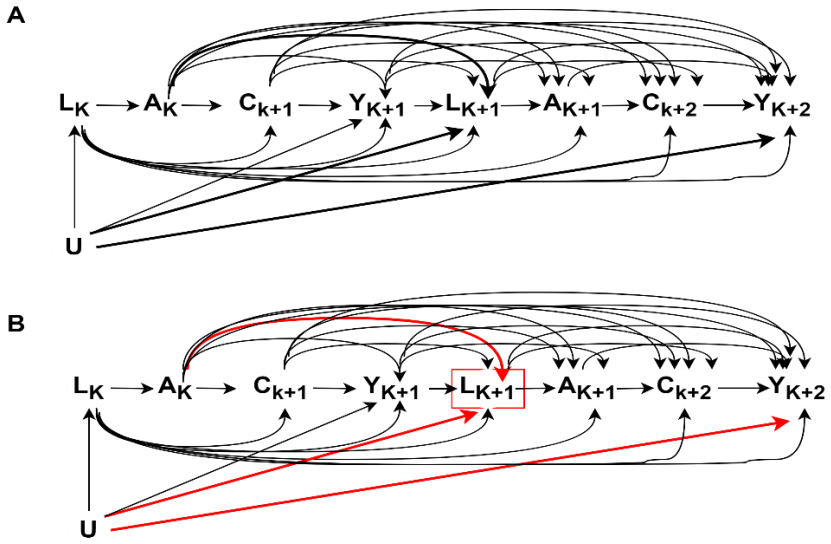

Figure 3. Causal directed acyclic graph depicted showing the structure of treatment-confounder feedback. Two follow-up periods are illustrated for simplicity: Panel (**A**) shows the data-generating structure where exchangeability holds, while Panel (**B**) highlights the bias (red arrows) introduced by conventional confounding adjustment methods. L_k: time-varying covariates at time k (e.g., body-mass index); A_k: time-varying exposures at time k (i.e., egg consumption); C_{k+1}: censoring at time $k+1$; Y_{k+1}: mortality at time $k+1$; U: unmeasured risk factors for time-varying covariates and mortality.

3.2 Specification of target trial

This analysis was designed to emulate a target trial (i.e., a hypothetical pragmatic trial that would have answered the causal question of interest) of intervention to reduce egg intake for the prevention of all-cause mortality. The key protocol components of this target trial are summarized in **Table 3**.

Table 3. Specification and emulation of a target trial of intervention to reduce egg consumption and all-cause mortality using observational data from the China Health and Nutrition Survey (1991-2015)

protocol component	target trial specification	target trial emulation
Eligibility criteria	Aged ≥20 years	The same criteria applied as for the target trial, with the following exceptions:

		➢ A response to the food frequency questionnaires from both 1991 and 1993 was required. ➢ Complete data was needed for both the baseline (1993) and pre-baseline (1991) periods on diet (with a plausible reported energy intake range of 800-4,200 kcal/day), age, and other risk factors.
Dietary strategy	No intervention (usual diet) Intervention to consume ≥1 serving of egg (≥50 gr) per day for non-vegetarians, and no intervention on egg intake for vegetarians.	The same criteria applied as for the target trial, with the following exceptions: ➢ The follow-up periods correspond to the waves in the cohort study. ➢ Each 3-year dietary intake assumed accurately reflects the individual's average dietary intake over the preceding 3-year period, as well as the intended diet that the individual would have reported at the start of the current period if no intervention had occurred.
Strategy assignment	Individuals are randomly assigned to a strategy at baseline and are aware of their assigned strategy.	Randomization was assumed to be conditional on the baseline covariates listed in the section above ('Covariates and confounding variables')
Outcomes	All-cause mortality	The same criteria applied as for the target trial. All-cause mortality was determined through questionnaires administered in each wave (approximately every 3 years), where household members of the deceased were asked to report the death.
Follow-up	For each eligible individual, the follow-up period begins at the time of strategy assignment (baseline) and continues until the individual is lost to follow-up, experiences the outcome of interest (death), or reaches December 2015, whichever occurs first.	The same criteria applied as for the target trial. Loss to follow-up was defined as the failure to respond to the questionnaire.
Causal contrasts	**Intention-to-treat and Per-protocol Effects:** The main focus of this study is to estimate the overall impact of the specified intervention strategy on all-cause mortality, considering all possible causal pathways between the intervention and outcomes.	Observational analogue of per-protocol effect

Statistical analysis	**Intention-to-treat Analysis:** This approach evaluates the effect based on the initial assignment of the intervention, regardless of adherence. **Per-protocol Analysis:** The parametric g-formula is used to compare the 22-year risk under each strategy, calculating risk ratios and risk differences. This analysis adjusts for both pre-baseline and post-baseline factors related to adherence and loss to follow-up.	The same approach was applied as in the per-protocol analysis of the target trial.

For each eligible individual, follow-up starts upon strategy assignment (baseline, 1993) and ends upon loss to follow-up, death (outcome of interest), or the administrative end of follow-up (December 2015), whichever happens first.

3.3 Emulation of the target trial

The observational data from the China Health and Nutrition Survey (CHNS), a long-term cohort study involving participants aged 20 to 101 years at the time of their enrolment in 1991, was used to replicate the target trial described above. Upon enrolment and roughly every three years thereafter, participants provided comprehensive information regarding medical diagnoses and lifestyle behaviors. Dietary intake data was similarly collected every three years. Deaths were identified through questionnaires administered to family members. Each element of the target trial was thoroughly replicated using this observational data (as detailed in **Table 3**). Values for egg consumption and other variables that changed over time were carried forward until updated by a subsequent questionnaire or until the participant was lost to follow-up (indicated by no response to the entire questionnaire). The key causal comparison was the observational counterpart of the per-protocol effect. Participants were assumed similar at baseline based on the covariates listed in **Supplementary Table 1** and adjusted for these baseline covariates to simulate randomization. The analysis mirrored the per-protocol analysis of the target trial.

3.4 Statistical analysis

The g-formula was used in this thesis (9). The g-formula, an extension of standardization for time-varying treatments and confounders, was used to estimate the 22-year risks of all-cause mortality associated with egg consumption. Unlike traditional stratification methods for adjusting confounding, the g-formula effectively addresses treatment-confounder feedback (**Figure 3**), where confounders such as BMI are influenced by earlier treatments like dietary interventions. This method has been applied in previous studies to estimate the impact of dietary interventions on hypertension, heart failure and mortality (39, 40). The risk of all-cause mortality was computed as if all participants had consistently followed the egg intake strategies throughout the entire follow-up period.

Cox proportional hazards models: In the analysis, an approach to measure egg consumption as exposure, similar to methods used in previous studies on single diet exposure and health outcomes was used. In this case, the baseline egg consumption was used. To compare the causal approach with traditional statistical methods, the most common approaches in the existing nutritional epidemiology literature, specifically the Cox proportional hazards model, was used (41). The hazard ratio obtained by the Cox model was then compared to the one obtained from the g-formula. Cox's model, commonly used in survival analysis, provides the "instantaneous failure rate" or hazard as a function of time. The hazard function, which represents the 'velocity' of event occurrence, has units of events per unit time and is always non-negative. In this model, exposure can be time-dependent or time-independent. The Proportional Hazards Model assumes that the effect of an exposure on time to event is consistent over time, multiplying the hazard function by a constant. This implies a constant hazard ratio (HR) for different exposure levels over time. However, hazards may not be proportional if treatment effects change over time or if disease susceptibility varies between individuals, leading to potential interaction or effect modification. When hazards are not proportional, the standard variance estimator from the Cox model may be incorrect, and the HR should be interpreted as a weighted average over the follow-up period (42).

Baseline characteristics (1993) was determined using descriptive statistics: proportion for categorical variables; mean for continuous and symmetrically distributed variables; and media for continuous and

asymmetrically distributed variables. All analysis were conducted with R software, version 4.3.1 (R Foundation for Statistical computing). A package called "gfoRmula" (43) was used to run the g-formula.

4. Results

The baseline characteristics of these men are shown in **Table 4**. In this study, the median age of participants was 42.6 years, with a nearly equal distribution between those consuming less than one serving of eggs per day and those consuming one or more servings. The proportion of males was slightly higher in the higher egg consumption group (50.7%) compared to the lower consumption group (47.8%). Urban residency was more common among higher egg consumers (45.8%) than lower consumers (30.6%). Those with higher egg consumption also had higher incomes, with 53.3% in the high-income bracket compared to 34.1% in the low consumption group. Education levels were higher in the higher egg consumption group, with 27.8% having a high level of education compared to 14.8% in the lower consumption group. Smoking status and alcohol consumption were similar across both groups, but BMI was slightly higher in the higher egg consumption group (median 22.2) compared to the lower consumption group (median 21.6). Physical activity levels were consistent across groups, and dietary factors showed differences, particularly in meat consumption, which was higher in the higher egg consumption group. The prevalence of hypertension was also slightly higher in the higher egg consumption group (14.7%) compared to the lower consumption group (11.5%). Between 1993 and 2015 (22 years), there were 776 (14.3%) deaths out of a total of 5,445 participants.

Table 4. Baseline characteristics of 5,445 eligible individuals when emulating a target trial of egg consumption and all-cause mortality, China Health and Nutrition Survey (1993)

		Egg consumption	
Characteristics	**Overall**	**<1 servings/day**	**≥1 serving/day**
N	**5,445**	**4739 (87.0%)**	**706 (13.0%)**
Age (years) – median (IQR)	42.6 [21.9, 93.4]	42.7 [21.9, 93.4]	42.1 [22.1, 82.6]
Sex (male)	2622 (48.2%)	2264 (47.8%)	358 (50.7%)
Residency (urban)	1773 (32.6%)	1450 (30.6%)	323 (45.8%)
Income			
Low	1612 (29.6%)	1508 (31.8%)	104 (14.7%)
Middel	1842 (33.8%)	1616 (34.1%)	226 (32.0%)
High	1991 (36.6%)	1615 (34.1%)	376 (53.3%)
Education			
Low	3041 (55.8%)	2743 (57.9%)	298 (42.2%)

Medium	1507 (27.7%)	1295 (27.3%)	212 (30.0%)
High	897 (16.5%)	701 (14.8%)	196 (27.8%)
Smoking			
Non-smoker	3520 (64.6%)	3063 (64.6%)	457 (64.7%)
Ex-smokers	98 (1.8%)	86 (1.8%)	12 (1.7%)
Current smokers	1827 (33.6%)	1590 (33.6%)	237 (33.6%)
Alcohol drink			
None	3420 (62.8%)	3002 (63.3%)	418 (59.2%)
<1/week	529 (9.7%)	468 (9.9%)	61 (8.6%)
1-2/week	532 (9.8%)	461 (9.7%)	71 (10.1%)
3-4/week	312 (5.7%)	254 (5.4%)	58 (8.2%)
Daily	652 (12.0%)	554 (11.7%)	98 (13.9%)
BMI – median (IQR)	21.6 [13.1, 38.9]	21.6 [13.4, 38.9]	22.2 [13.1, 34.0]
Physical activity (MET-hours/week) – Median (IQR)	200 [0.0375, 1750]	200 [0.0375, 1500]	200 [0.263, 1750]
Dietary factors (grams/day) - median (IQR)			
Egg	0 [0, 1050]	0 [0, 48.4]	66.6 [50.0, 1050]
Vegetables and fruit	383 [0, 2450]	392 [0, 2450]	333 [0, 2390]
Legumes and nuts	0 [0, 528]	0 [0, 528]	0 [0, 267]
Whole grain	0 [0, 972]	0 [0, 972]	0 [0, 417]
Meat	50.0 [0, 732]	50.0 [0, 732]	71.1 [0, 458]
Fish	0 [0, 556]	0 [0, 556]	0 [0, 258]
Energy intake (kcal) – mean (SD)	2280 (756)	2290 (761)	2250 (723)
Hypertension	648 (11.9%)	544 (11.5%)	104 (14.7%)

BMI – body mass index; IQR – interquartile range

Figure 4 displays the cumulative incidence of death over time, categorized by two levels of egg consumption: less than 50 grams per day ("<50 gm/day") and 50 grams or more per day (">=50 gm/day"). The x-axis represents the waves (years of follow-up), and the y-axis shows the cumulative incidence of death.

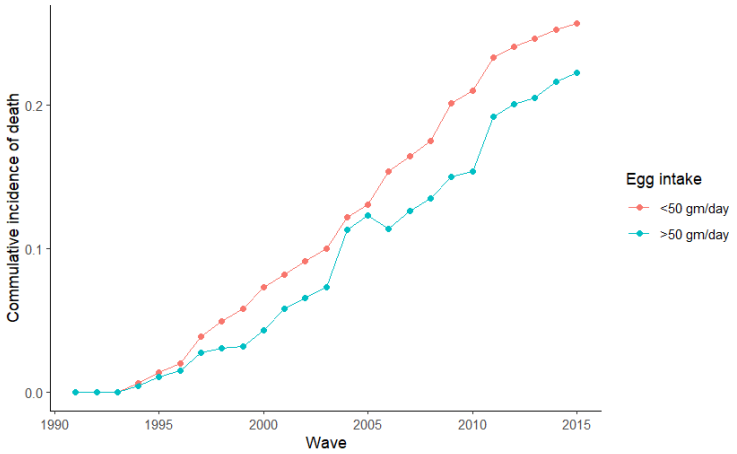

Figure 4. Cumulative death by egg consumption

In this 22-year follow-up study, the application of the g-formula revealed that individuals consuming less than one egg per day had an estimated risk of 10% (95% CI: 9% to 11%) for all-cause mortality. In contrast, the intervention group, where individuals consumed one or more eggs per day, demonstrated a slightly reduced risk of 9.1% (95% CI: 8.4% to 9.4%). The absolute risk difference between the two groups was -1% (95% CI: -1.7% to -0.6%), indicating a modest but statistically significant risk reduction associated with higher egg consumption. The relative risk ratio for the intervention was 0.902 (95% CI: 0.83 to 0.93), suggesting an approximate 9.8% reduction in risk compared to the control group. Nearly all individuals (99.9%) in the intervention group were successfully intervened upon, with a cumulative intervention rate of 67.5% over the study period. These findings suggest that higher egg consumption may be associated with a slight reduction in long-term risk for all-cause mortality studied (**Table 5**).

In contrast, the Cox proportional hazards model suggests a similar trend with a hazard ratio of 0.91, but with a wider confidence interval (95% CI: 0.69, 1.20), indicating that the reduction in hazard is imprecise. While both methods suggest a potential benefit to reduce all-cause mortality from higher egg consumption, the g-formula provides a more precise estimate of the risk reduction (**Table 5**).

Table 5. Estimated 22-year risks of all-cause mortality under sustained egg consumption using g-formula and comparison with Cox model, China Health and Nutrition Survey (1990-2016)

	22-year risk (%) (95% CI)	**Risk difference (%) (95% CI)**	**Risk ratio (95% CI)**	**Average % intervened on[a]**	**Cumulative % intervened on[b]**
G-formula					
No intervention[c] (<1 egg/day)	0.10 (0.09, 0.11)	-	1.00	-	-
Intervention (≥1 egg/day)	0.091 (0.084, 0.094)	-0.010 (-0.017, -0.006)	0.90 (0.83, 0.93)	99.9%	67.5%
Cox-model[d]			**Hazard ratios (95% CI)**		
No intervention (<1 egg/day)			1.00		
Intervention (≥1 egg/day)			0.91 (0.69, 1.20)		

[a] Average proportion of participants who would have to change their egg consumption in each 3-year period to keep adhering to the specified strategy.
[b] Cumulative proportion of participants who would have to change their egg consumption in any time period over follow-up to keep adhering to the specified strategy.
[c] No intervention corresponds to the "natural course" (observed egg consumption in this population).
[d] A hazard ratio from a Cox model needs to be interpreted as a weighted average of the true HRs over the entire follow-up period. But, it is wrongly interpreted as the treatment effect during the follow-up period.

Figure 5 illustrates that over a 21-year period, individuals consuming one egg or more per day consistently experience a lower cumulative risk compared to those consuming less than one egg per day. By year 21, the cumulative risk for the higher egg consumption group is around 9%, while it exceeds 10% for the lower consumption group, suggesting a modest protective effect associated with higher egg intake.

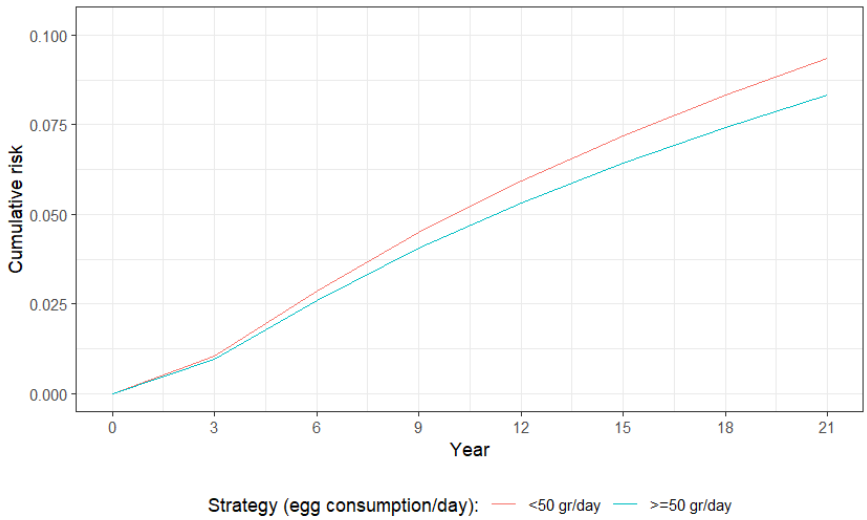

Figure 5. Estimated survival under hypothetical egg intake interventions compared with no intervention in the China Health and Nutrition Survey. Estimates are based on the parametric g-formula with pre-baseline, baseline and time-varying covariates

5. Discussion

5.1 Summary of findings

In this 22-year follow-up study, the g-formula analysis revealed that individuals consuming less than one egg per day had an increased risk for all-cause mortality. In contrast, those consuming one or more eggs per day showed a slightly reduced risk. This indicates a modest but statistically significant risk reduction associated with higher egg consumption. The relative risk ratio of 0.902 (95% CI: 0.83 to 0.93) suggests a 9.8% reduction in risk compared to the lower consumption group. The intervention was nearly universally successful, with 99.9% of participants adhering to the higher egg

consumption strategy, leading to a cumulative intervention rate of 67.5% over the study period. In contrast, the Cox proportional hazards model indicated a similar trend but with a less precise hazard ratio, suggesting an imprecise estimate of risk reduction. Both methods support the potential benefit of higher egg consumption, though the g-formula provides a more precise estimate of risk reduction. Overall, these findings suggest that higher egg consumption is associated with a slight reduction in long-term risk for the outcome studied. The g-formula's results, along with the supporting evidence from the Cox model, underscore the potential protective effect of sustained higher egg intake.

5.2 Comparison with other studies

In this 22-year follow-up study, individuals consuming less than one egg per day had a 10% risk of the outcome of interest, while those consuming one or more eggs per day demonstrated a slightly reduced risk of 9.1%. The absolute risk difference of -1% indicates a modest but statistically significant reduction in risk associated with higher egg consumption. This finding aligns with several previous studies that have explored the relationship between egg consumption and all-cause mortality, though the magnitude and direction of the association have varied. For example, Djoussé et al. (22) reported a higher relative risk of 1.23 (95% CI: 1.11, 1.36) for individuals consuming seven or more eggs per week compared to those consuming less than one per week, suggesting an increased risk associated with high egg consumption. Similarly, Zhong et al. (29) found a relative risk of 1.25 (95% CI: 1.04, 1.52) for individuals consuming two or more eggs per day compared to those who consumed none, further indicating a potential increase in risk with higher egg intake.

However, not all studies have reported an increased risk. Mann et al. (19) observed a relative risk of 0.92 (95% CI: 0.68, 1.23) for individuals consuming more than six eggs per week compared to those consuming less than one egg per week, suggesting no significant association between egg consumption and mortality. This discrepancy could be attributed to differences in study populations, dietary assessment methods, and the confounding factors accounted for in each study. For instance, Virtanen et al. (27) reported a protective effect of higher egg consumption, with a relative risk of 0.74 (95% CI: 0.63, 0.87) for individuals consuming 59 grams of eggs per day compared to those consuming only 8 grams per day. This finding supports the potential benefit of moderate egg consumption in reducing long-term mortality risk.

Our study's findings, particularly the precise estimates provided by the g-formula, suggest a potential modest benefit from higher egg consumption, contrasting with some previous studies that indicated increased risk. The discrepancy in results across studies underscores the complexity of dietary research and the need for careful consideration of study design, population characteristics, and confounding factors. The broader literature, including the findings of Farvid et al. (25), who observed a relative risk of 0.88 (95% CI: 0.79, 0.97) for those consuming 26.4 grams of eggs per day compared to none, suggests that the impact of egg consumption on mortality may be context-dependent, influenced by overall diet quality, lifestyle factors, and individual health conditions. Therefore, while our study adds to the evidence base, further research is necessary to fully elucidate the relationship between egg consumption and long-term health outcomes.

5.3 Limitations of the study

This study has several limitations that should be considered when interpreting the results. Firstly, biomarkers were not taken into account, which limits the precision of our findings related to biological processes underlying the observed associations. Secondly, the analysis of all-cause mortality was based on crude data, which includes deaths from various causes, including accidents, that may not be directly related to dietary factors. This approach may introduce noise into our mortality estimates, reducing their specificity. Furthermore, no sensitivity analyses were conducted to assess the impact of potential confounders, which could have provided a more robust understanding of the results. As with any observational study, we cannot entirely eliminate the possibility of unmeasured or residual confounding. Despite our efforts to adjust for numerous potential confounders, there remains a risk that unmeasured factors, such as health care-seeking behaviors, might influence our findings. For instance, individuals who are more health-conscious might be more inclined to adhere to healthier diets, engage more with healthcare services, and therefore, might have a higher likelihood of being diagnosed with a specific cause of death, which could bias the results.

The study also relied on self-reported dietary information, which is inherently subject to measurement error. Although the repeated 24-hour recall method used in this study provides a useful estimate of average dietary intake, it is still prone to inaccuracies and recall bias. Additionally, our findings may not be generalizable to other populations, particularly those with different dietary patterns, risk factors,

or other effect modifiers. Finally, it should be noted that missing data were not imputed in our analysis. The exclusion of participants with missing data could potentially lead to biased results, particularly if the missing data are not random. This limitation further highlights the need for cautious interpretation of our findings.

5.4 Public health implications of the findings and future directions

The findings of this study have significant public health implications and suggest several avenues for future research. The findings from this study suggest that higher egg consumption may be associated with a modest reduction in long-term mortality risk. This highlights the potential role of eggs as part of a balanced diet in promoting longevity. However, given the variability in results across different studies, public health recommendations should emphasize overall dietary quality and individual health conditions when advising on egg consumption. Further research is needed to confirm these findings and inform more specific dietary guidelines.

One potential direction is to conduct substitution analysis with other dietary components, such as meat, to better understand the impact of different food groups on health outcomes. This could provide insights into healthier dietary choices and guide public health recommendations. To address residual confounding, sensitivity analyses using negative controls, such as deaths due to accidents, could be performed. This would help to determine if the associations observed in the study are truly related to dietary factors or if they are influenced by other, unrelated factors. It may also be beneficial to exclude participants who had pre-existing diseases at baseline, as these individuals might have already modified their dietary behaviour in response to their health conditions. This exclusion would provide a clearer picture of the impact of diet on disease development in otherwise healthy individuals.

Further analysis could focus on younger populations by removing older participants, who may have different health behaviours and risks. Analysing data from younger participants could offer more relevant insights for public health interventions aimed at preventing disease in earlier stages of life. Moreover, future studies could adjust for covariates from the previous wave rather than using data from the questionnaire concurrent with dietary information. This approach would help in understanding the temporal relationships between diet and health outcomes more accurately. Additionally, specifying

different functional forms for the covariates could allow for a more nuanced analysis and help identify non-linear relationships. Finally, to mitigate possible reverse causation, participants who died within the first two years of follow-up could be excluded from the analysis. This would reduce the likelihood that the observed associations are driven by individuals who were already in poor health at the start of the study and whose dietary behaviours might have been influenced by their deteriorating condition. These future directions could enhance the robustness of the findings and provide more precise recommendations for dietary interventions in public health.

6. Conclusion

In conclusion, this 22-year follow-up study provides evidence that higher egg consumption is associated with a modest reduction in long-term mortality risk. The application of the g-formula offered a precise estimate of this risk reduction, which was supported, albeit less definitively, by the Cox proportional hazards model. These findings contribute to the ongoing debate about the health implications of egg consumption, suggesting that, when consumed as part of a balanced diet, eggs may have a protective effect on longevity. However, the variability in results across different studies underscores the importance of considering individual dietary patterns and health conditions in public health recommendations. Further research is essential to solidify these conclusions and guide future dietary guidelines. Diet is a highly complex exposure that interacts with an array of social, behavioural, physiological, and genetic factors. Building causal knowledge in nutritional epidemiology requires the triangulation of evidence from diverse areas such as nutritional biochemistry, metabolism, the food ecosystem, and other multifaceted factors beyond nutrition. This complexity underscores the fact that short-term RCTs, while invaluable for revealing the short-term effects of diet on intermediate health markers and for elucidating pathophysiological and molecular mechanisms, may not be sufficient to generate comprehensive evidence. To better understand diet's long-term impacts on health and disease, it is crucial to enrich and triangulate the findings from RCTs with evidence gathered from meticulously designed and executed long-term observational studies. However, resolving current inconsistencies in nutritional epidemiology evidence cannot be achieved by merely replacing observational studies with RCTs. Instead, refining current methods, developing new data collection

and analysis tools, and harnessing emerging technologies in nutritional science will enhance the robustness, consistency, and overall value of evidence from this field.

7. References

1. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1923-94.
2. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2019;393(10184):1958-72.
3. Auestad N, Fulgoni VL, 3rd. What current literature tells us about sustainable diets: emerging research linking dietary patterns, environmental sustainability, and economics. Adv Nutr. 2015;6(1):19-36.
4. Grosso G, Laudisio D, Frias-Toral E, Barrea L, Muscogiuri G, Savastano S, et al. Anti-Inflammatory Nutrients and Obesity-Associated Metabolic-Inflammation: State of the Art and Future Direction. Nutrients. 2022;14(6).
5. Wang YB, Page AJ, Gill TK, Melaku YA. The association between diet quality, plant-based diets, systemic inflammation, and mortality risk: findings from NHANES. Eur J Nutr. 2023;62(7):2723-37.
6. Zhao J, Li Z, Gao Q, Zhao H, Chen S, Huang L, et al. A review of statistical methods for dietary pattern analysis. Nutr J. 2021;20(1):37.
7. Stern D, Ibsen DB, MacDonald CJ, Chiu YH, Lajous M, Tobias DK. Improving nutrition science begins with asking better questions. Am J Epidemiol. 2024.
8. Lanza ST, Coffman DL, Xu S. Causal Inference in Latent Class Analysis. Struct Equ Modeling. 2013;20(3):361-83.
9. Hernán MA, Robins JM (2020). Causal inference: what if. Boca Raton: Chapman & Hall/CRC.
10. Stoetzer LF, Zhou X, Steenbergen M. Causal inference with latent outcomes. American Journal of Political Science.n/a(n/a).
11. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med. 2014;33(12):2137-55.
12. Global Dietary Database , 2020. Accessed 05 April 2024. Available: https://www.globaldietarydatabase.org/.
13. USDA: United States Department of Agriculture. Choice Reviews Online. 2011;48(07):48-3859–48-3859. doi: 10.5860/CHOICE.48-3859.
14. Li MY, Chen JH, Chen C, Kang YN. Association between Egg Consumption and Cholesterol Concentration: A Systematic Review and Meta-analysis of Randomized Controlled Trials. Nutrients. 2020;12(7).
15. Rouhani MH, Rashidi-Pourfard N, Salehi-Abargouei A, Karimi M, Haghighatdoost F. Effects of Egg Consumption on Blood Lipids: A Systematic Review and Meta-Analysis of Randomized Clinical Trials. J Am Coll Nutr. 2018;37(2):99-110.
16. Godos J, Micek A, Brzostek T, Toledo E, Iacoviello L, Astrup A, et al. Egg consumption and cardiovascular risk: a dose-response meta-analysis of prospective cohort studies. Eur J Nutr. 2021;60(4):1833-62.
17. Drouin-Chartier JP, Schwab AL, Chen S, Li Y, Sacks FM, Rosner B, et al. Egg consumption and risk of type 2 diabetes: findings from 3 large US cohort studies of men and women and a systematic review and meta-analysis of prospective cohort studies. Am J Clin Nutr. 2020;112(3):619-30.

18. Kahn HA, Phillips RL, Snowdon DA, Choi W. Association between reported diet and all-cause mortality: twenty-one-year follow-up on 27, 530 adult Seventh-Day Adventists. American Journal of Epidemiology. 1984;119(5):775-87.
19. Mann JI, Appleby PN, Key TJ, Thorogood M. Dietary determinants of ischaemic heart disease in health conscious individuals. Heart. 1997;78(5):450-5.
20. Nakamura Y, Okamura T, Tamaki S, Kadowaki T, Hayakawa T, Kita Y, et al. Egg consumption, serum cholesterol, and cause-specific and all-cause mortality: the National Integrated Project for Prospective Observation of Non-communicable Disease and Its Trends in the Aged, 1980 (NIPPON DATA80). The American journal of clinical nutrition. 2004;80(1):58-63.
21. Qureshi AI, Suri MFK, Ahmed S, Nasar A, Divani AA, Kirmani JF. Regular egg consumption does not increase the risk of stroke and cardiovascular diseases. Medical Science Monitor. 2006;13(1):CR1-CR8.
22. Richman EL, Kenfield SA, Stampfer MJ, Giovannucci EL, Chan JM. Egg, red meat, and poultry intake and risk of lethal prostate cancer in the prostate-specific antigen-era: incidence and survival. Cancer prevention research. 2011;4(12):2110-21.
23. Wang J-B, Fan J-H, Dawsey SM, Sinha R, Freedman ND, Taylor PR, et al. Dietary components and risk of total, cancer and cardiovascular disease mortality in the Linxian Nutrition Intervention Trials cohort in China. Scientific reports. 2016;6(1):22619.
24. Yu D, Zhang X, Xiang Y-B, Yang G, Li H, Gao Y-T, et al. Adherence to dietary guidelines and mortality: a report from prospective cohort studies of 134,000 Chinese adults in urban Shanghai. The American journal of clinical nutrition. 2014;100(2):693-700.
25. Farvid MS, Malekshah AF, Pourshams A, Poustchi H, Sepanlou SG, Sharafkhah M, et al. Dietary protein sources and all-cause and cause-specific mortality: the Golestan Cohort Study in Iran. American journal of preventive medicine. 2017;52(2):237-48.
26. Nakamura Y, Okamura T, Kita Y, Okuda N, Kadota A, Miura K, et al. Re-evaluation of the associations of egg intake with serum total cholesterol and cause-specific and total mortality in Japanese women. European Journal of Clinical Nutrition. 2018;72(6):841-7.
27. Virtanen HE, Voutilainen S, Koskinen TT, Mursu J, Kokko P, Ylilauri MP, et al. Dietary proteins and protein sources and risk of death: the Kuopio Ischaemic Heart Disease Risk Factor Study. The American journal of clinical nutrition. 2019;109(5):1462-71.
28. Chen G-C, Chen L-H, Mossavar-Rahmani Y, Kamensky V, Shadyab AH, Haring B, et al. Dietary cholesterol and egg intake in relation to incident cardiovascular disease and all-cause and cause-specific mortality in postmenopausal women. The American journal of clinical nutrition. 2021;113(4):948-59.
29. Zhong VW, Van Horn L, Cornelis MC, Wilkins JT, Ning H, Carnethon MR, et al. Associations of dietary cholesterol or egg consumption with incident cardiovascular disease and mortality. Jama. 2019;321(11):1081-95.
30. Zupo R, Sardone R, Donghia R, Castellana F, Lampignano L, Bortone I, et al. Traditional dietary patterns and risk of mortality in a longitudinal cohort of the Salus in Apulia Study. Nutrients. 2020;12(4):1070.

31. Ruggiero E, Di Castelnuovo A, Costanzo S, Persichillo M, De Curtis A, Cerletti C, et al. Egg consumption and risk of all-cause and cause-specific mortality in an Italian adult population. European journal of nutrition. 2021;60:3691-702.
32. Satija A, Yu E, Willett WC, Hu FB. Understanding nutritional epidemiology and its role in policy. Adv Nutr. 2015;6(1):5-18.
33. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. Jama. 2022;328(24):2446-7.
34. Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. Bmj. 2022;378:e071108.
35. Zhang B, Zhai FY, Du SF, Popkin BM. The China Health and Nutrition Survey, 1989-2011. Obes Rev. 2014;15 Suppl 1(0 1):2-7.
36. Yu D, Zhao L, Zhang J, Yang Z, Yang L, Huang J, et al. China Nutrition and Health Surveys (1982-2017). China CDC Wkly. 2021;3(9):193-5.
37. Popkin BM, Du S, Zhai F, Zhang B. Cohort Profile: The China Health and Nutrition Survey--monitoring and understanding socio-economic and health change in China, 1989-2011. Int J Epidemiol. 2010;39(6):1435-40.
38. Zhai FY, Du SF, Wang ZH, Zhang JG, Du WW, Popkin BM. Dynamics of the Chinese diet and the role of urbanicity, 1991-2011. Obes Rev. 2014;15 Suppl 1(0 1):16-26.
39. Chiu YH, Chavarro JE, Dickerman BA, Manson JE, Mukamal KJ, Rexrode KM, et al. Estimating the effect of nutritional interventions using observational data: the American Heart Association's 2020 Dietary Goals and mortality. Am J Clin Nutr. 2021;114(2):690-703.
40. Ibsen DB, Chiu YH, Gémes K, Wolk A. Hypothetical 22-Year Intervention With the Dietary Approaches to Stop Hypertension and Risk of Heart Failure in a General Population. Am J Epidemiol. 2024;193(1):96-106.
41. Collett, D. (2014). Modelling Survival Data in Medical Research (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18041.
42. Hernán MA. The hazards of hazard ratios. Epidemiology. 2010;21(1):13-5.
43. McGrath S, Lin V, Zhang Z, Petito LC, Logan RW, Hernán MA, et al. gfoRmula: An R Package for Estimating the Effects of Sustained Treatment Strategies via the Parametric g-formula. Patterns (N Y). 2020;1(3).

Appendix

Supplementary Table 1. Variables used to model 22-year risk of all-cause mortality, China Health and Nutrition Survey (1993-2015).			
	Years assessed	**As dependent variable**	**As independent variable**
Time-fixed (baseline) variables			
Age	1993	Not predicted	Continuous
Sex	1993	Not predicted	Indicator: 0 - male 1 - female
Residency	1993	Not predicted	Indicator 0 – rural 1 – urban
Income	1993	Not predicted	3 categories 0 – low 1 – middle 2 – high
Education	1993	Not predicted	3 categories 0 – low 1 – medium 2 – high
Smoking	1993	Not predicted	3 Categories 0 - non-smoker 1 - ex-smokers 2 - current smokers
Alcohol drink	1993	Not predicted	5 Categories 0 - None 1 - <1/week 2 - 1-2/week

Supplementary Table 1. Variables used to model 22-year risk of all-cause mortality, China Health and Nutrition Survey (1993-2015).			
	Years assessed	**As dependent variable**	**As independent variable**
			3 - 3-4/week 4 - Daily
Body mass index	1993	Not predicted	Continuous
Hypertension	1993	Not predicted	Indicator: 0. No 1. Yes
Cardiovascular disease	1993	Not predicted	Indicator: 0. No 1. Yes
Diabetes	1993	Not predicted	Indicator: 0. No 1. Yes
Time-fixed (pre-baseline) variables			
Eggs	1991	Not predicted	Indicator: 0 - <1 servings/day 1 - ≥1 serving/day
Vegetables and fruit	1991	Not predicted	Continuous
Legumes and nuts	1991	Not predicted	Continuous
Whole grain	1991	Not predicted	Continuous
Meat	1991	Not predicted	Continuous
Fish	1991	Not predicted	Continuous
Soft drinks	1991	Not predicted	Continuous
Total energy intake	1991	Not predicted	Continuous
Body mass index (BMI)	1991	Not predicted	Continuous
Time-varying variables:			

Supplementary Table 1. Variables used to model 22-year risk of all-cause mortality, China Health and Nutrition Survey (1993-2015).			
	Years assessed	**As dependent variable**	**As independent variable**
When used as independent variables, values in the current and the previous period were included in the models (except for follow-up time). Modeling the joint distribution of time-varying covariates reported in waves.			
Follow-up period	N/A	Not predicted	7 categories (1 for each 3-year period) for all models
Income	1991-2011, every three years	Logistic then log-linear	3 categories 0 – low 1 – middle 2 – high
Education	1991-2011, every three years	Logistic then log-linear	3 categories 0 – low 1 – medium 2 – high
Smoking	1991-2011, every three years	Logistic then log-linear	3 Categories 0 - non-smoker 1 - ex-smokers 2 - current smokers
Alcohol drink	1991-2011, every three years	Logistic then log-linear	5 Categories 0 - None 1 - <1/week 2 - 1-2/week 3 - 3-4/week 4 - Daily
Body mass index	1991-2011, every three years	Linear	Continuous
Hypertension	1991-2011, every three years	Logistic to failure	Indicator: 0. No 1. Yes

Supplementary Table 1. Variables used to model 22-year risk of all-cause mortality, China Health and Nutrition Survey (1993-2015).			
	Years assessed	**As dependent variable**	**As independent variable**
Cardiovascular disease	1991-2011, every three years	Logistic to failure	Indicator: 0. No 1. Yes
Diabetes	1991-2011, every three years	Logistic to failure	Indicator: 0. No 1. Yes
Eggs	1991-2011, every three years	Logistic regression	Indicator: 0 - <1 servings/day 1 - ≥1 serving/day
Vegetables and fruit	1991-2011, every three years	Logistic then log-linear	Continuous
Legumes and nuts	1991-2011, every three years	Logistic then log-linear	Continuous
Whole grain	1991-2011, every three years	Logistic then log-linear	Continuous
Meat	1991-2011, every three years	Logistic then log-linear	Continuous
Fish	1991-2011, every three years	Logistic then log-linear	Continuous
Total energy intake	1991-2011, every three years	Logistic then log-linear	Continuous
Body mass index (BMI)	1991-2011, every three years	Logistic then log-linear	Continuous

R code

DATA PREPARATION ADN DISCRIPTIVE ANALYSIS

#1 - Import the data

```r, results=FALSE}
#| warning: false
library(tidyverse); library(tidyr); library(VIM); library(lubridate); library(ggsci)

```

Variables to consider as confounders: SES: age, sex, urban, income (income3c), urban3c, education (edu3c) Behavioral: smoking (smoke), alcohol drinking (drink), exercise (pa) Metabolic : hypertension(hbp), bmi Disease: cvd, diabetes (kndm) and cancer (no data)

Food items: vegetable, fruit, *whole_grain*, legumes, nuts, *meat*, poultry, *fish*, milk, wheat_rice, beverage \[*vege_fruit*, *legumes_nuts*\] Energy intake: total energy intake (day_kcal) Outcome: death; death time: dyear other variables we need: idind, hhid, commid, line

```r}
library(haven)
d_egg <- read_stata("egg_mortality_data.dta")
table (d_egg$death)
dim(d_egg)
table(d_egg$kndm, d_egg$wave)
table(d_egg$cvd, d_egg$wave)
#d_egg <- d_egg |> mutate(across(where(is.labelled), as.numeric))


#Select those variables that we are interested in
d_egg <- d_egg |>
    dplyr::select (idind, hhid, commid, line, wave, age, urban, urban3c,
                   smoke, pa, drink, hbp, bmi,
                   cvd, kndm,
                   death, ddate,
                   egg, whole_grain, meat,
                   fish, vege_fruit, legumes_nuts, sex,
                   urban, income3c, edu3c, smoke, drink,
                  day_kcal)
```

```r}
#create new id variable
d_egg <- d_egg |>
        group_by(idind) |>
            mutate(id = cur_group_id()) |>
            ungroup()
#create new household variable
d_egg <- d_egg |>
        group_by(hhid) |>
            mutate(idhh = cur_group_id()) |>
            ungroup()
dim(d_egg)
length(unique(d_egg$id))
colnames(d_egg)

```

##1.1 - Missing values - pre-baseline (1991) & baseline (1993)

```r}
table (d_egg$wave)

#prebaseline data (1991)
pre_BL.data <- d_egg |> subset(wave==1991)
pre_BL.data$kndm[is.na(pre_BL.data$kndm)] <- 0
pre_BL.data$cvd[is.na(pre_BL.data$cvd)] <- 0
missing_values <- sapply(pre_BL.data, function(x) sum(is.na(x)))
missing_values
complete_PBL.data <- na.omit(pre_BL.data)
dim(complete_PBL.data)
missing_values1 <- sapply(complete_PBL.data, function(x) sum(is.na(x)))
missing_values1
dim(complete_PBL.data)


#Baseline data (1993)
BL.data <- d_egg |> subset(wave==1993)
BL.data$kndm[is.na(BL.data$kndm)] <- 0
BL.data$cvd[is.na(BL.data$cvd)] <- 0
missing_values <- sapply(BL.data, function(x) sum(is.na(x)))
missing_values
complete_BL.data <- na.omit(BL.data)
```

```
dim(complete_BL.data)
missing_values1 <- sapply(complete_BL.data, function(x) sum(is.na(x)))
missing_values1
dim(complete_BL.data)

#Merge 2 datasets
PBL_BL.data <- merge(complete_BL.data, complete_PBL.data, by = "idind", all.x = TRUE)
dim(PBL_BL.data)


#those participated in both 1991 and 1993 (pre-baseline and baseline)
PBL_BL.id <- PBL_BL.data |> select(idind)
dim(PBL_BL.id)


egg_data <- merge(d_egg, PBL_BL.id, by = "idind", all.x = FALSE)
dim(egg_data)
egg_data <- egg_data[!is.na(egg_data$wave), ]

library(dplyr)
egg_data <- egg_data |> arrange(idind, wave)



#cvd and diabetes are missing at 1991 and 1993 (no cases)
length(unique(egg_data$idind))
egg_data$cvd[egg_data$wave %in% c(1991, 1993) & is.na(egg_data$cvd)] <- 0
egg_data$kndm[egg_data$kndm %in% c(1991, 1993) & is.na(egg_data$kndm)] <- 0


#We have to remove anyone who didn't participated in 1991 and 1993 but participated in the next
waves
# Step 1: Identify idind values that appear in 1993 but not in 1991
idind_1991 <- unique(egg_data$idind[egg_data$wave == 1991])
idind_1993_not_1991 <- unique(egg_data$idind[egg_data$wave == 1993 & !egg_data$idind %in%
idind_1991])

# Step 2: Remove observations with those idind values from the 1993 wave
egg_data <- egg_data[!(egg_data$wave == 1993 & egg_data$idind %in% idind_1993_not_1991), ]


length(unique(egg_data$idind))
```

```
table(egg_data$wave)


#We have to remove anyone who didn't participated in 1991 and 1993 but participated in the next waves

# Step 1: Identify idind values that did not participate in 1991 and 1993
idind_not_1991_1993 <- unique(egg_data$idind[!(egg_data$idind %in% egg_data$idind[egg_data$wave %in% c(1991, 1993)])])

# Step 2: Identify idind values that participated in 1997, 2000, 2004, 2006, 2009, 2011
idind_participated_later <- unique(egg_data$idind[egg_data$wave %in% c(1997, 2000, 2004, 2006, 2009, 2011)])

# Step 3: Find idind values that did not participate in 1997 and 2000 but did participate later
idind_to_remove <- intersect(idind_not_1991_1993, idind_participated_later)

# Step 4: Remove those participants from egg_data
egg_data <- egg_data[!(egg_data$idind %in% idind_to_remove), ]

length(unique(egg_data$idind))

table(egg_data$death)

#ids of the participants for death  merge
id.data <- egg_data |> select(idind, wave)

egg_data <- egg_data |> select(-death, -ddate)
table(egg_data$wave)
```


#2 - Death dataset

```r}
#Prepare death dataset by selecting those with death variable
d_egg_death <- d_egg |>
      dplyr::select(idind, id, wave, death, ddate)

missing_counts <- d_egg_death |>
            summarise(across(everything(), ~ sum(is.na(.))))
missing_counts
```

```
#Remove those with death missing values  - date of death is missing for 7 people
d_egg_death <-  d_egg_death  |>
            drop_na(death, ddate) |>
            filter(death == 1)

# Convert the 'ddate' column to a Date object if it's not already
d_egg_death$ddate <- as.Date(d_egg_death$ddate)

# Extract the year and create a new column
d_egg_death$dyear <- year(d_egg_death$ddate)

# View the updated dataset
table(d_egg_death$dyear)



table(d_egg_death$dyear)
d_egg_death <- d_egg_death |> filter(dyear>1993)
d_egg_death <- d_egg_death |> select(idind, death, dyear)


#Merge the mortality data with id from the data
merged_data <- merge(d_egg_death, id.data, by = "idind", all.x = TRUE)
table(merged_data$death)
table(merged_data$death, merged_data$wave)


death.data <- na.omit(merged_data)
table (death.data$wave)


death.data <- death.data |> select(idind, death, dyear)
death.data <- death.data |> distinct(idind, .keep_all = TRUE)

dim(death.data)

#Merge death data with original data (clean and complete)
# Perform the merge
egg.all.data <- merge(egg_data, death.data, by = "idind", all.x = TRUE)
table(egg.all.data$wave, egg.all.data$death)
dim(egg.all.data)
```

```
egg.all.data.2 <- egg.all.data

#Make death=1 only for the latest wave
# First, order the data by idind and wave to ensure the latest wave is at the bottom for each idind
#

# Step 1: Sort the data by idind and wave
    #arrange the variables
    egg.all.data.2 <- egg.all.data.2 |>
                select(idind, wave, sex, age, death, dyear, edu3c, income3c, urban,
                    smoke, pa, drink, bmi, hbp, cvd, kndm, egg, whole_grain, meat,
                    fish, vege_fruit, legumes_nuts, day_kcal)
    #sort by idind and wave
    egg.all.data.2 <- egg.all.data.2 |>
                arrange(idind, wave)
          #replace death=0 if NA
          egg.all.data.2$death[is.na(egg.all.data.2$death)] <- 0
          length(is.na(egg.all.data.2$death))
          table(egg.all.data.2$death)
          dim(egg.all.data.2)

#remove observations where edu3c is NA, death = 1, and the wave is the last one for that idind
egg.all.data.2 <- egg.all.data.2 %>%
 arrange(idind, wave) %>%  # Ensure data is sorted by idind and wave
 group_by(idind) %>%
 filter(!(is.na(edu3c) & death == 1 & wave == max(wave))) %>%
 ungroup()

   table(egg.all.data.2$wave, egg.all.data.2$death, useNA = "ifany")

egg.all.data.2$kndm[is.na(egg.all.data.2$kndm) & egg.all.data.2$wave==1991] <- 0
egg.all.data.2$kndm[is.na(egg.all.data.2$kndm) & egg.all.data.2$wave==1993] <- 0



# Step 2: Assign death = 1 to the last wave
egg.all.data.2 <- egg.all.data.2 %>%
      group_by(idind) %>%
      mutate(death = ifelse(death == 1 & wave == max(wave[death == 1]), 1, 0)) %>%
      ungroup()
table(egg.all.data.2$wave, egg.all.data.2$death)
```

```
dim(egg.all.data.2)

#remove participants entirely if they have death = 1 at either wave 1991 or 1993.

        #     egg.all.data.3 <- egg.all.data.2 %>%
        #     group_by(idind) %>%
        #     filter(!any(death == 1 & wave %in% c(1991, 1993))) %>%
        #     ungroup()
        #
        #      table(egg.all.data.3$wave)
        #      table(egg.all.data.3$wave, egg.all.data.3$death, useNA = "ifany")
        # #Egg category
        #         egg.all.data.3 <- egg.all.data.3 %>%
        #           mutate(egg_cat = ifelse(egg < 50, 0, 1))
        #
        #         table(egg.all.data.3$wave, useNA = "ifany")
        #



#Replace the previous wave if the current wave is missing
colnames(egg.all.data)
egg.all.data.3 <- egg.all.data.2 %>%
 arrange(idind, wave) %>%    # Ensure the data is sorted by idind and wave
 group_by(idind) %>%         # Group by idind to perform operations within each individual
 fill(sex, edu3c, income3c, urban, smoke, pa, drink, bmi, hbp, cvd, kndm, egg, whole_grain, meat,
fish, vege_fruit, legumes_nuts, day_kcal, .direction = "down") %>%  # Replace NA with the previous
non-NA value
 ungroup()  # Ungroup after the operation

table(egg.all.data.3$wave)
length(is.na(egg.all.data.3$edu3c))
table(egg.all.data.2$death, egg.all.data.2$wave, useNA = "ifany")



# Define the variables to check for missing values
variables_to_check <- c("idind", "wave", "sex", "age", "death", "edu3c", "income3c",
                        "urban", "smoke", "pa", "drink", "bmi", "hbp", "cvd",
                        "kndm", "egg", "whole_grain", "meat", "fish",
                        "vege_fruit", "legumes_nuts", "day_kcal")
```

```
# Generate the miss variable
egg.all.data.4 <- egg.all.data.3 %>%
  mutate(miss = ifelse(if_any(all_of(variables_to_check), is.na), 1, 0))
table(egg.all.data.4$miss, egg.all.data.4$death)


#remove all observations across all waves for any idind where miss = 1 specifically in wave 1991
# Step 1: Generate the miss variable for wave 1991
egg.all.data.4 <- egg.all.data.4 %>%
  mutate(miss = ifelse(wave == 1991 & if_any(all_of(variables_to_check), is.na), 1, 0))

# Step 2: Identify idind with miss = 1 in wave 1991
idind_to_remove <- egg.all.data.4 %>%
  filter(wave == 1991 & miss == 1) %>%
  pull(idind) %>%
  unique()

# Step 3: Remove all observations for those idind across all waves
egg.all.data.4 <- egg.all.data.4 %>%
  filter(!idind %in% idind_to_remove)
table(egg.all.data.4$wave, egg.all.data.4$death)
table(egg.all.data.4$wave, egg.all.data.4$death)


egg.all.data.4 <- egg.all.data.4 %>%
  mutate(egg_cat = ifelse(egg < 50, 0, 1))

```


#3 - Table 1

```r}


bl.data <- egg.all.data.4  |>
   dplyr:::filter(wave == 1993) |>
   select(sex, age, death, edu3c, income3c,
                  urban, smoke, pa, drink, bmi, hbp, cvd,
                  kndm, egg, egg_cat, whole_grain, meat, fish,
                  vege_fruit, legumes_nuts, day_kcal)
library(table1)
```

```
# label(bl.data$sex) <- "Sex"
# label(bl.data$age) <- "Age (years)"
# label(bl.data$egg) <- "Egg (gm/d)"
# label(bl.data$fish) <- "Fish (gm/d)"
# label(bl.data$meat) <- "Meat (gm/d)"
# label(bl.data$death_final) <- "Death"

bl.data$sex <- as.factor(bl.data$sex)
bl.data$urban <- as.factor(bl.data$urban)
bl.data$income3c <- as.factor(bl.data$income3c)
bl.data$edu3c <- as.factor(bl.data$edu3c)
bl.data$smoke <- as.factor(bl.data$smoke)
bl.data$drink <- as.factor(bl.data$drink)
bl.data$hbp <- as.factor(bl.data$hbp)
bl.data$cvd<- as.factor(bl.data$cvd)
bl.data$death <- as.factor(bl.data$death)

summary_table <- table1(~ age + sex + urban + income3c + edu3c + smoke + drink + bmi + pa +
                          egg + vege_fruit + legumes_nuts + whole_grain + meat + fish +
                          day_kcal + hbp + cvd + kndm + death | egg_cat,
                        data = bl.data)

# Convert the summary table to a data frame for export
# Note: The output of table1 is usually more complex (formatted for display),
# so you may need to extract or manually create the content as a data frame.
summary_table_df <- as.data.frame(summary_table)

# Export the summary table to a CSV file
write.csv(summary_table_df, "summary_table_by_egg_cat.csv", row.names = FALSE)


save(bl.data, egg.all.data.4, file = "my_data.RData")

```

#4 - Figure mortality by egg

```r}
# Prepare the data
data <- egg.all.data.4 %>%
  dplyr::select(idind, wave, death, egg) %>%
```

```
  dplyr::mutate(egg_serve = ifelse(egg < 50, "<50 gm/day", ">=50 gm/day"))

# Calculate cumulative deaths
data <- data %>%
 arrange(idind, wave) %>%  # Ensure the data is sorted by idind and wave
 group_by(idind) %>%
 mutate(cumulative_death = cumsum(death)) %>%  # Calculate cumulative deaths for each
participant
 ungroup()

# Calculate cumulative incidence for each wave and egg_serve category
data_proportions <- data %>%
 group_by(wave, egg_serve) %>%
 summarize(cumulative_incidence = mean(cumulative_death > 0, na.rm = TRUE), .groups = "drop")
%>%
 ungroup()

# Plot the cumulative incidence
ggplot(data_proportions, aes(x = wave, y = cumulative_incidence, group = egg_serve, color =
as.factor(egg_serve))) +
 geom_point() +
 geom_line() +
 theme_classic() +
 labs(x = "Wave", y = "Cumulative Incidence", title = "Cumulative Incidence of Death by Egg
Consumption", color = "gm/day")


```
```

MODELLING

#1 - Data
```r}
load("my_data.RData")
```

#2 -Data preparation for g-formula

```r}
data <- egg.all.data.4
table(data$wave)
```
```

```
data$sex <- as.factor(data$sex)
data$urban <- as.factor(data$urban)
data$income3c <- as.factor(data$income3c)
data$edu3c <- as.factor(data$edu3c)
data$smoke <- as.factor(data$smoke)
data$drink <- as.factor(data$drink)
data$hbp <- as.numeric(data$hbp)
data$cvd<- as.factor(data$cvd)
data$death <- as.numeric(data$death)
data$egg_cat <- as.numeric(data$egg_cat)


#replace waves 1991, 1993, 1997, 2000, 2004, 2006, 2009, 2011 to 0, 3, 6, 9, 12, 15, 18, 21

data <- data %>%
 mutate(Year = recode(wave,
                `1991` = 0,
                `1993` = 3,
                `1997` = 6,
                `2000` = 9,
                `2004` = 12,
                `2006` = 15,
                `2009` = 18,
                `2011` = 21))


table(data$Year)

data.1 <- data %>%
 arrange(idind, Year)

# Function to remove participants who participated in a given year but not in the previous year
remove_participants <- function(data.1, year1, year2)
 idind_to_remove <- data.1 %>%
  filter(Year == year1) %>%
  anti_join(data.1 %>% filter(Year == year2), by = "idind") %>%
  pull(idind)

 data.1 <- data.1 %>%
  filter(!idind %in% idind_to_remove)
```

```
  return(data.1)
}

# Apply the function for each pair of years
data.1 <- data.1 %>%
  remove_participants(21, 18) %>%  # Remove participants who participated in Year 21 but not in
Year 18
  remove_participants(15, 12) %>%  # Remove participants who participated in Year 15 but not in
Year 12
  remove_participants(9, 6) %>%    # Remove participants who participated in Year 9 but not in Year 6
  remove_participants(3, 0)        # Remove participants who participated in Year 3 but not in Year 0


# Apply the function for each pair of years
data.1 <- data.1 %>%
  remove_participants(21, 0) %>%
  remove_participants(18, 0) %>%
  remove_participants(15, 0) %>%
remove_participants(12, 0) %>%
  remove_participants(9, 0) %>%
    remove_participants(6, 0) %>%
  remove_participants(3, 0)

data.1 <- data.1 %>%
  remove_participants(21, 3) %>%
  remove_participants(18, 3) %>%
  remove_participants(15, 3) %>%
remove_participants(12, 3) %>%
  remove_participants(9, 3) %>%
    remove_participants(6,3)

table(data.1$Year)
table(data.1$Year, data.1$death)

#Creat a variable to show how many times a participant participated across years
data.1 <- data.1 %>%
  arrange(idind, wave) %>%  # Sort data by participant ID and wave
  group_by(idind) %>%  # Group by participant ID
  mutate(participated = row_number()) %>%  # Create the cumulative participation count
  ungroup()

table(data.1$participated, data.1$Year)
```

```
year_correct <- c(0, 3, 6, 9, 12, 15, 18, 21)

# Create the correct variable
data.1 <- data.1 %>%
 mutate(correct = ifelse(Year == year_correct[participated], 1, 0))


table(data.1$Year, data.1$death)
table(data.1$death)

temp.data <- data.1 |> filter(correct==0)
  View(temp.data)


```

#3 - Censoring

```r
# Function to assign censored status
assign_censored <- function(data.1, current_year, previous_year)
 # Identify participants who are not in the current year but were in the previous year
 data.1 <- data.1 %>%
  group_by(idind) %>%
  mutate(
   censored = case_when(
    Year == previous_year & !idind %in% data$idind[data$Year == current_year] & death == 0 ~ 1,
    Year == previous_year & !idind %in% data$idind[data$Year == current_year] & death == 1 ~ 0,
    TRUE ~ censored
   )
  ) %>%
  ungroup()

 return(data.1)
}

# Initialize censored variable to 0
data.1$censored <- 0

# Apply the function for each pair of years
```

```
data.1 <- data.1 %>%
  assign_censored(6, 3) %>%   # Check between Year 6 and Year 3
  assign_censored(9, 6) %>%   # Check between Year 9 and Year 6
  assign_censored(12, 9) %>%  # Check between Year 12 and Year 9
  assign_censored(15, 12) %>% # Check between Year 15 and Year 12
  assign_censored(18, 15) %>% # Check between Year 18 and Year 15
  assign_censored(21, 18)     # Check between Year 21 and Year 18

table(data.1$censored, data.1$death, useNA = "ifany")
table(data.1$censored, data.1$participated, useNA = "ifany")



data.1 <- data.1 %>%
  mutate(time = recode(Year,
                       '0'=0,
                       '3'=1,
                       '6'=2,
                       '9'=3,
                       '12'=4,
                       '15'=5,
                       '18'=6,
                       '21'=7))

table(data.1$time)

```


#4 - Modelling
##4.1 - data preparation

```r}
# load required packages
library(gfoRmula)
library(Hmisc)


#Pre-baseline variables
data.1$time <- as.numeric(data.1$time)
pre_bl.data <- data.1 |> dplyr::filter(time==0)
```

```
# Add 'pre_' prefix to all variables except 'idind'
pre_bl.data <- pre_bl.data %>%
  rename_with(~ paste0("pre_", .), -idind)


# Add 'pre_' prefix to all variables except 'idind'
bl.data <- data.1 |> dplyr::filter(time==1)
bl.data <- bl.data %>%
  rename_with(~ paste0("bl_", .), -idind)
View(bl.data)

# Merge data.1 with pre_bl.data
merged_data <- data.1 %>%
  left_join(pre_bl.data, by = "idind") %>%  # Keep all rows from data.1
  left_join(bl.data, by = "idind")          # Keep all rows from data.1
View(merged_data)

table(merged_data$time)


save(bl.data, egg.all.data.4, merged_data, file = "my_data.RData")

setDT(merged_data)
# Ensure the data is sorted by id and time
# Assuming the correct column names are idind and time
if ("idind" %in% colnames(merged_data) & "time" %in% colnames(merged_data))
  # Ensure the data is sorted by idind and time
  merged_data <- merged_data %>%
    arrange(idind, time)

  # Create a full sequence of time points for each idind
  full_time <- data.table(expand.grid(idind = unique(merged_data$idind), time = 0:8))

  # Merge with the original data
  merged_data <- full_time[merged_data, on = .(idind, time)]

  # Fill missing values by carrying the last observation forward
  merged_data <- merged_data %>%
    group_by(idind) %>%
    fill(-time, .direction = "down") %>%
    ungroup()
```

```
 # Convert back to data.table if needed
 setDT(merged_data)

 # View the resulting merged data
 head(merged_data)
} else
 stop("Column names 'idind' and/or 'time' not found in merged_data")
}

merged_data <- merged_data |> filter(correct==1)
table(merged_data$time, merged_data$participated)


merged_data$egg_cat <- as.numeric(merged_data$egg_cat)


#
# merged_data <- merged_data %>%
#   arrange(idind, time) %>%
#   group_by(idind) %>%
#   mutate(new_time = row_number() - 1) %>%
#   ungroup()
#
# table(merged_data$time, merged_data$participated, useNA = 'ifany')

```


##4.2 - g-formula

```r}

#data <- merged_data |> filter(wave!=1991)
data <- merged_data
# table(data$death, data$time)

# data <- data %>%
#   mutate(time = recode(Year,
#                        '3'=0,
#                        '6'=1,
#                        '9'=2,
#                        '12'=3,


```

```
#                 '15'=4,
#                 '18'=5,
#                 '21'=6))


class(data$edu3c)
class(data$income3c)
class(data$smoke)
class(data$drink)


data$death <- as.numeric(data$death)
unique(data$death)
table(data$death, data$wave, useNA='ifany')
class(data$death)
data <- data[data$death %in% c(0, 1), ]

data <- data.table(merged_data)
data$kndm <- as.numeric(data$kndm)
table(data$kndm)

data$egg_cat <- as.numeric(data$egg_cat)


data$hbp <- as.numeric(data$hbp)
#overall setup
#Specifying parameters for implementing parametric g-formula

setDT(data)
seed = 1246
id <- "idind"
length(unique(data$time))
time_points <- 8
time_name <- "time"
nsimul <- 10000
library(parallel)
ncores <- as.numeric(parallel::detectCores())
parallel::detectCores()

#exposure
intvars <- list('egg_cat', 'egg_cat')
```

```
interventions <- list(list(c(static, rep(0, time_points))),
                      list(c(static, rep(1, time_points))))

int_descript <- c('<50 gram/day', '>=50 gram/day')

#outome
outcome_type <- "survival"

outcome_name <- "death"



ymodel <- death ~ age + sex + urban +
                  income3c + edu3c +
                  smoke + pa + drink +
                  bmi + hbp +
                  lag1_whole_grain + lag1_vege_fruit +
                  lag1_fish + lag1_legumes_nuts + lag1_meat +
                  lag1_egg_cat + lag1_day_kcal +
                  Hmisc::rcspline.eval(vege_fruit, nk=3) +
                  whole_grain + vege_fruit + fish +
                  legumes_nuts + meat + egg_cat + day_kcal  + time

#covariates
basecovs <- c("age", "sex", "urban")

covnames <- c("income3c", "edu3c",
              "smoke", "pa", "drink",
              "bmi", "hbp",
              "vege_fruit","whole_grain" ,
              "fish" , "legumes_nuts" ,
              "meat" , "egg_cat" ,
              "day_kcal")

data$hbp <- as.numeric(data$hbp)
# Check levels of all categorical variables
sapply(data[, covnames, with = FALSE], function(x) if(is.factor(x)) levels(x))

# Check for variables with only one level
sapply(data[, covnames, with = FALSE], function(x) if(is.factor(x)) length(unique(x)))
```

```
covtypes <- c("categorical", "categorical",
              "categorical", "normal", "categorical",
              "normal", "binary",
              "zero-inflated normal", "zero-inflated normal",
              "zero-inflated normal","zero-inflated normal",
              "zero-inflated normal","binary",
              "normal")

histories=c(lagged)

histvars <- list(c(
              "income3c", "edu3c",
              "smoke", "pa", "drink",
              "bmi", "hbp",
              "whole_grain" , "vege_fruit" ,
              "fish" , "legumes_nuts" ,
              "meat" , "egg_cat" ,
              "day_kcal"))

# Covariate models


covparams <- list(covmodels = c(
  income3c ~ age + sex + urban +
    lag1_edu3c + edu3c +
    lag1_income3c +
    lag1_smoke + smoke +
    lag1_pa +  pa +
    lag1_drink + drink +
    lag1_bmi + bmi +
    lag1_hbp + hbp +
    #lag1_kndm + kndm +
    #lag1_cvd + cvd +
    lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
    lag1_whole_grain + whole_grain +
    lag1_fish + fish +
    lag1_legumes_nuts +  legumes_nuts +
    lag1_meat + meat +
    lag1_egg_cat + egg_cat  +
```

```
  lag1_day_kcal  + day_kcal +
  time,
 edu3c ~ age + sex + urban +
  lag1_edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd + cvd +
  lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
 smoke ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd + cvd +
  lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
pa ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c  +
  lag1_smoke + smoke +
```

```
  lag1_pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd + cvd +
  lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
drink ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd + cvd +
  lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
bmi ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd + cvd +
```

```
  lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
hbp ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp +
  # lag1_kndm + kndm +
  # lag1_cvd + cvd +
  lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
 # kndm ~ age + sex + urban +
 #   lag1_edu3c + edu3c +
 #   lag1_income3c + income3c +
 #   lag1_smoke + smoke +
 #   lag1_pa +  pa +
 #   lag1_drink + drink +
 #   lag1_bmi + bmi +
 #   lag1_hbp + hbp +
 #   lag1_kndm +
 #   lag1_cvd +  cvd +
 #   lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
 #   lag1_whole_grain + whole_grain +
 #   lag1_fish + fish +
 #   lag1_legumes_nuts +  legumes_nuts +
 #   lag1_meat + meat +
 #   lag1_egg_cat + egg_cat  +
```

```
#   lag1_day_kcal  + day_kcal +
#   time,
# cvd ~ age + sex + urban +
#   lag1_edu3c + edu3c +
#   lag1_income3c + income3c +
#   lag1_smoke + smoke +
#   lag1_pa +  pa +
#   lag1_drink + drink +
#   lag1_bmi + bmi +
#   lag1_hbp + hbp +
#   lag1_kndm + kndm +
#   lag1_cvd +
#   lag1_vege_fruit + Hmisc::rcspline.eval(vege_fruit, nk=3) +
#   lag1_whole_grain + whole_grain +
#   lag1_fish + fish +
#   lag1_legumes_nuts +  legumes_nuts +
#   lag1_meat + meat +
#   lag1_egg_cat + egg_cat  +
#   lag1_day_kcal  + day_kcal +
#   time,
vege_fruit ~ age + sex + urban +
 lag1_edu3c + edu3c +
 lag1_income3c + income3c +
 lag1_smoke + smoke +
 lag1_pa +  pa +
 lag1_drink + drink +
 lag1_bmi + bmi +
 lag1_hbp + hbp +
 # lag1_kndm + kndm +
 # lag1_cvd +  cvd +
 lag1_vege_fruit +
 lag1_whole_grain + whole_grain +
 lag1_fish + fish +
 lag1_legumes_nuts +  legumes_nuts +
 lag1_meat + meat +
 lag1_egg_cat + egg_cat  +
 lag1_day_kcal  + day_kcal +
 time,
whole_grain ~ age + sex + urban +
 lag1_edu3c + edu3c +
 lag1_income3c + income3c +
 lag1_smoke + smoke +
```

```
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd +  cvd +
  lag1_vege_fruit +  Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
fish~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd +  cvd +
  lag1_vege_fruit +  Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish +
  lag1_legumes_nuts +  legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
legumes_nuts ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd +  cvd +
```

```
  lag1_vege_fruit +  Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
meat ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd +  cvd +
  lag1_vege_fruit +  Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts + legumes_nuts +
  lag1_meat +
  lag1_egg_cat + egg_cat  +
  lag1_day_kcal  + day_kcal +
  time,
egg_cat ~ age + sex + urban +
  lag1_edu3c + edu3c +
  lag1_income3c + income3c +
  lag1_smoke + smoke +
  lag1_pa +  pa +
  lag1_drink + drink +
  lag1_bmi + bmi +
  lag1_hbp + hbp +
  # lag1_kndm + kndm +
  # lag1_cvd +  cvd +
  lag1_vege_fruit +  Hmisc::rcspline.eval(vege_fruit, nk=3) +
  lag1_whole_grain + whole_grain +
  lag1_fish + fish +
  lag1_legumes_nuts + legumes_nuts +
  lag1_meat + meat +
  lag1_egg_cat +
```

```
    lag1_day_kcal  + day_kcal +
    time,
  day_kcal ~ age + sex + urban +
    lag1_edu3c + edu3c +
    lag1_income3c + income3c +
    lag1_smoke + smoke +
    lag1_pa +  pa +
    lag1_drink + drink +
    lag1_bmi + bmi +
    lag1_hbp + hbp +
    # lag1_kndm + kndm +
    # lag1_cvd +  cvd +
    lag1_vege_fruit +  Hmisc::rcspline.eval(vege_fruit, nk=3) +
    lag1_whole_grain + whole_grain +
    lag1_fish + fish +
    lag1_legumes_nuts + legumes_nuts +
    lag1_meat + meat +
    lag1_egg_cat + egg_cat +
    lag1_day_kcal  +
    time))


# Parametric g-formula
# with competing risks - but not estimated using IPW
results <- gformula(
            obs_data = data,
            id = id,
            outcome_name = outcome_name,
            outcome_type = outcome_type,
            ymodel = ymodel,
            time_name = time_name,
            covnames = covnames,
            covparams = covparams,
            covtypes = covtypes,
            basecovs = basecovs,
            intvars = intvars,
            interventions = interventions,
            histories = histories,
            histvars = histvars,
            nsimul = nsimul,
            parallel = TRUE,
            nsamples = 5,
```

```
    ncores = ncores,
    seed = 1234
  )


main_results <- results$result
print(main_results)

main_results_df <- as.data.frame(main_results)
#install.packages("writexl")
library(writexl)
write_xlsx(main_results_df, "main_results.xlsx")

# e) Risk curves
j <- 3 # Total number of interventions (2 + natural course)
k <- 8 # 48 times points

expand.grid(k=0:(k), Interv.=1:(j-1)) %>%
 left_join(data.frame(results$result) %>% mutate(k=k+1), by=c("k", "Interv.")) %>%
 group_by(Interv.) %>%
 mutate(
  g.form.risk = case_when(
   k == 0 ~ 0,
   TRUE ~ g.form.risk),
  Intervention = case_when(
   Interv. == 1 ~ "<50 gr/day",
   Interv. == 2 ~ ">=50 gr/day")) %>%
 mutate(g.form.risk = g.form.risk) %>%
 ggplot(aes(x = k * 3, y = g.form.risk, color = Intervention)) +  # Multiply k by 3
 geom_line() +
 scale_x_continuous("Year", limits = c(0, 21), breaks = seq(0, 21, 3)) +
 scale_y_continuous("Cumulative risk") +
 theme_bw() +
 labs(colour = "Strategy (egg consumption/day):") +
 theme(legend.position = "bottom")


```


#5. Survivala anlsysis

```r}
# Load necessary packages
#install.packages("survival")
library(survival)

write.csv(data, "data.csv")



last_wave_data <- data %>%
  group_by(idind) %>%              # Group by individual identifier
  filter(time == max(time)) %>%     # Keep only the row(s) with the maximum time value
  ungroup()                        # Ungroup the data

surv_obj <- Surv(time = last_wave_data$wave, event = last_wave_data$death)



last_wave_data$time_death <- last_wave_data$dyear - 1991
last_wave_data$Year <- ifelse(!is.na(last_wave_data$time_death),
                              last_wave_data$time_death,
                              last_wave_data$Year)

# Increase iterations
cox_model <- coxph(Surv(Year, death) ~ pre_egg_cat + pre_age + pre_sex + pre_urban +
pre_income3c + pre_edu3c +
                     pre_smoke + pre_pa + pre_drink + pre_bmi + pre_hbp + pre_vege_fruit +
                     pre_whole_grain + pre_fish + pre_legumes_nuts + pre_meat ,
                     data = last_wave_data)

summary(cox_model)


```