A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease

Soodabeh Milanlouei, Giulia Menichetti, Yanping Li, Joseph Loscalzo, Walter C. Willett & Albert-László Barabási

Environmental factors, and in particular diet, are known to play a key role in the development of Coronary Heart Disease. Many of these factors were unveiled by detailed nutritional epidemiology studies, focusing on the role of a single nutrient or food at a time. Here, we apply an Environment-Wide Association Study approach to Nurses' Health Study data to explore comprehensively and agnostically the association of 257 nutrients and 117 foods with coronary heart disease risk (acute myocardial infarction and fatal coronary heart disease). After accounting for multiple testing, we identify 16 food items and 37 nutrients that show statistically significant association - while adjusting for potential confounding and control variables such as physical activity, smoking, calorie intake, and medication use - among which 38 associations were validated in Nurses’ Health Study II. Our implementation of Environment-Wide Association Study successfully reproduces prior knowledge of diet-coronary heart disease associations in the epidemiological literature, and helps us detect new associations that were only marginally studied, opening potential avenues for further extensive experimental validation. We also show that Environment-Wide Association Study allows us to identify a bipartite food-nutrient network, highlighting which foods drive the associations of specific nutrients with coronary heart disease risk.
The prevalence of heart disease, the leading cause of death throughout the world, is strongly influenced by diet and eating habits\(^1\). For example, a recent CDC (Centers for Disease Control and Prevention) report\(^3\), focusing on death rates caused by heart disease across the United States, documented substantial regional differences compatible with different eating patterns. Similarly, while among individuals of Japanese descent, coronary heart disease (CHD) incidence rates are only 1.6 per person-years in Japan, it increases to 3.0 in Hawaii, and 3.7 in San Francisco\(^5\), differences that cannot be explained by genetic factors, documenting the key role dietary and other environmental factors play in the development of the disease.

Much of our knowledge about the role of food on health comes from epidemiological association studies in which a single or limited number of exposure(s) is/are analyzed in relation to a phenotype, representing a hypothesis-driven path towards understanding diet–disease relationships. Yet, diet is not a simple sum of several nutrients, as each food product consists of a mixture of nutrients associated with multiple compounds of limited or unknown nutritional value\(^6\). Accordingly, the effect of each dietary compound on human health should not be investigated in isolation, but in the presence of other associated chemical compounds and relevant food sources. For example, Kolonel et al.\(^8\) initially reported that beta-carotene consumption was positively associated with the risk of prostate cancer. While this finding was worrisome, in a subsequent analysis of foods, intake of carrots, the largest source of beta-carotene, was not related to the risk of prostate cancer; the observed association was due to intake of papaya\(^9\). Thus, the analysis of foods provided evidence against the effect of beta-carotene and suggested that some factors specific to papaya might be responsible for the original finding\(^10\).

An alternative approach is dietary pattern analysis that focuses on the effects of the overall diet\(^11\)–\(^13\), rather than a single or a few nutrients. While dietary pattern analyses are ideal in the development of nutritional guidelines\(^14\), they are insufficient for the agnostic discovery of new signals for further experimental or mechanistic validation.

As an alternative to the traditional epidemiological studies, environment-wide association studies (or EWAS) were proposed to identify new environmental factors in disease and disease-related phenotypes in an unbiased manner. EWAS is inspired by the analytical procedures developed in genome-wide association studies (GWAS)\(^15\) in which a panel of “exposures” (genotype variants) is studied in relation to a phenotype of interest. For example, using the National Health and Nutrition Examination Survey dataset, an EWAS study explored the associations of 543 environmental attributes with type 2 diabetes, identifying five statistically significant associations validated across independent cohorts\(^15\). Wulaningsih et al.\(^16\) investigated 182 nutrition and lifestyle factors in relation to abdominal obesity, finding a statistically significant association of obesity with five factors in men and seven factors in women. Merritt et al.\(^17\) used European Prospective Investigation into Cancer and Nutrition (EPIC) data to evaluate endometrial cancer risk associations for the dietary intake of 84 foods and nutrients, concluding that only coffee intake had a statistically significant inverse relationship.

Despite the recent success of the EWAS methodology in unveiling multiple nutritional factors that together may contribute to our health, its widespread use is undermined by several factors and limitations\(^18\). Indeed, failing to achieve adequate statistical power in association detection, EWAS studies could not always recover known environment–disease associations confirmed by large, prospective cohort studies and randomized trials\(^17\)\(^\text{-}19\)\(^,\)\(^20\). As we show below, these failures are not inherent in the EWAS methodology, but are mainly rooted in the limited size, limited variability, and lack of repeated measurements of the datasets to which EWAS has been applied thus far. While the statistical power of the EWAS study approach is a legitimate concern, the magnitude of the statistical power depends on multiple factors, including the nature and the size of the dataset, as well as the statistical tools/models used for the analysis. Indeed, as we show here, if we apply a wide-association study approach to an adequately sized longitudinal cohort dataset with sufficient variability, we consistently recover prior knowledge about diet–disease relationships.

The EWAS methodology may be particularly useful for diseases for which nutritional associations are unknown\(^18\). While the effect of dietary exposures on heart disease has been extensively studied and the causal effects of many of these associations confirmed, the diet–disease literature occasionally demonstrates conflicting findings\(^21\)–\(^23\), limiting our understanding of the true effect of dietary exposures on diseases. We will show that the wide-association study approach can provide comparable insights in an efficient manner by applying an unbiased standardized set of analytical tools.

Here we implement an EWAS methodology, aiming to identify dietary factors associated with CHD systematically and comprehensively, focusing on both nutrient intake and food consumption. To overcome the limited statistical power of previous studies, we apply our methodology to the Nurses’ Health Study (NHS), a longitudinal prospective study designed to investigate the longitudinal effects of nutrition on health and disease development. While there are larger cohorts available for studying chronic diseases, such as the UK Biobank\(^24\) and the China Kadoorie Biobank\(^25\), both with around 500,000 participants, NHS is unique, owing to comprehensive longitudinal dietary data collection. Beginning in 1976, NHS gathered registered female nurses, ages 30–55 years, from across the United States, initially designed to investigate the use of oral contraception in relation with risk of breast cancer. Participants are asked to complete questionnaires every 2 years, and in 1980 a Food Frequency Questionnaire (FFQ) has been included, designed to capture dietary behaviors. Follow-up dietary questionnaires were administered in 1984, 1986, and every 4 years since then. Questionnaires used from 1984 and thereafter included about 130 foods plus detailed information about brands and types of margarine, breakfast cereals, multiple vitamins, and types of fat used for cooking and baking. As health professionals, nurses were chosen for their ability to complete the health-related questionnaires thoroughly and accurately\(^26\). To date, NHS has been expanded to NHS II and NHS III to cover a younger population of nurses\(^27\). These three cohorts resulted in an extensive published body of research on the relationships of environmental and genetic factors to various diseases\(^28\). The dietary drivers of CHD have been extensively studied within NHS data, most analyses primarily focusing on a single or limited number of exposures, while controlling for an appropriate set of adjusting variables. Some of these findings, as those for trans-fats, have inspired experimental studies and were confirmed to have a causal effect on developing heart diseases\(^29\)–\(^31\).

To have a broad picture of the existing knowledge about diet–disease associations in the NHS data, we mined the literature to identify all studies exploring the dietary determinants of heart-related diseases in original NHS and successive cohorts. The resulting knowledge graph (Fig. 1) shows that the most extensively studied cardiovascular phenotype using NHS data is CHD. Here, we use the term negative when a higher level of exposure is associated with a lower CHD risk. Similarly, we use positive term when a higher level of an exposure is associated with CHD risk. We made this choice for simplicity, and it should not be confused with a causal relationship.
Excluding studies of biomarkers and tissues, the neighborhood of CHD comprises nutrients, food items, and dietary scores. Cumulatively, 120 associations were studied, documenting 63 negative associations and 22 positive relations with CHD; in the remaining cases, there was no significant association between dietary exposure and CHD risk, as in the case of dietary magnesium. The space of studied exposures is rather heterogeneous and is often driven by either the researcher’s interests or evidence from animal or mechanistic studies.

As illustrated in the knowledge graph, the single-association studies using NHS data have broadened our understanding of the dietary determinants of CHD. Some non-significant associations were found to be significant after the application of new statistical approaches in larger datasets. For example, while a study using the original NHS data found no association between fruit fiber intake and risk of CHD, a pooled analysis of three NHS cohorts found that the higher consumption of dietary fiber from fruits was negatively associated with risk of CHD.

Here, we show that a wide-association approach allows us to scan efficiently and systematically the dietary determinants of CHD, bypassing the problem with missing significant associations in epidemiological studies. By applying EWAS methodology to the NHS data, we find that a wide-association approach not only recovers the existing knowledge on diet–disease association, but also facilitates the discovery of novel associations, potentially inspiring future follow-up studies.

Fig. 1 The Knowledge Graph of the Dietary Factors Associated with Cardiovascular Disease in the NHS Data. The nodes of the graph are dietary exposures (circles) and cardiovascular diseases (diamonds) as two sets of nodes. Each studied association is shown by an edge whose color denotes the “direction” (i.e., positive or negative) of the association. Green and red links indicate negative and positive associations, respectively, between an exposure and a disease. Grey links denote associations that were studied but not found to be statistically significant. In the context of NHS, CHD refers to non-fatal MI and fatal coronary heart disease; also, coronary artery disease (CAD) refers to non-fatal MI and fatal coronary artery disease. Cardiovascular disease (CVD) is defined as a composite of coronary artery disease and non-fatal or fatal stroke. Source data are provided in Source Data - Figure 1.xlsx. The figure aims to illustrate the body of work derived from NHS data on cardiovascular diseases. A detailed comparison of the result of our analysis with previous NHS-related work and other findings in the literature is provided in the section “Comparison with the literature”.

Results
Main findings. During the follow-up period, 2774 incident cases of non-fatal MI or fatal CHD were documented in NHS. The baseline average of total caloric intake and body mass index (BMI) among participants who later developed CHD were slightly higher than in those who did not develop CHD. In addition, prospective case subjects on average had lower physical activity compared with the non-case population (Supplementary Table 1). We examined the effect of 374 exposures on CHD risk, including 257 nutrients and 117 food items. The descriptive characteristics of these exposures are shown in Supplementary Table 2.

For the first phase of EWAS, using Cox regression models we collected the estimated effect size, the variance of effect size, and hazard ratio (HR) for one standard deviation, P value, the 95% confidence interval (CI) for the HR, P value regarding the proportionality assumption, and the variance inflation factor (VIF). Consider, for example, the results of the long-term effect of isorhamnetin—an O-methylated flavonol—from the class of flavonoids (Fig. 2b). The estimated HR (0.91, 95% CI: 0.87–0.95; P value 1.59 × 10−5) implies that one standard deviation higher consumption of Box–Cox-transformed isorhamnetin is associated with 91% lower CHD risk. The P value regarding the proportionality assumption indicates that the use of the Cox model is appropriate. The VIF equal to 1.27 suggests that there is no severe multicollinearity among the variables involved in the isorhamnetin test.
Figure 3a shows the distribution of Cox model $P$ values for all investigated exposures. Exposures are ordered by the estimated HR, so that exposures with HR > 1.0 have harmful effects on CHD risk and exposures with HR < 1.0 are expected to be beneficial. Insignificant associations are mainly distributed around an HR of one, and exposures with smaller $P$ values are scattered at the two ends of the distribution. Using the permutation procedure to account for multiple testing, we found 53 significant associations, including 16 food items and 37 nutrients. All significant associations had VIF < 5. For all significant associations, except for phytate, the $P$ value regarding the proportionality assumption is >0.05. The minimum statistical power for detecting the smallest absolute effect size was 0.59, which is considered to be a moderate to a high level of power in clinical studies. A list of exposures that have a statistically significant association with CHD risk, together with their estimated HRs, is shown in Table 1. A list of both significant and non-significant associations is provided in Supplementary Table 3.

Figure 2 Ewas Methodology Description and Output. a For each exposure ($i$), we fit a Cox model to estimate CHD risk, while controlling for a set of adjusting variables ($n$ exposures, in total). Through this process, we also collect the VIF and the $P$ value regarding the proportionality assumption. After the fitting phase (yellow), we proceed with the multiple testing protocols (blue). We leverage confounding variables to estimate the likelihood of CHD development for each subject, and the disease status is then permuted accordingly, $m$ times. For each exposure, the FDR is estimated, and associations with an FDR < 0.05 are considered statistically significant. b Isorhamnetin is an $O$-methylated flavonol from the class of flavonoids mainly found in green pepper, red onion, and dill. EWAS shows that isorhamnetin intake is negatively associated with CHD risk (HR: 0.91; 95% CI: 0.87–0.95; $P$ value: $1.59 \times 10^{-5}$, from two-sided Wald test, with no adjustment for multiple comparisons). The $P$ value regarding the proportionality assumption indicates the appropriateness of using the Cox model. The VIF of 1.27 is an indication of the absence of severe multicollinearity.
Traditional epidemiological studies are limited to the detection of a single exposure (food or nutrient) in relation to CHD; however, the developed EWAS methodology allows us to explore the space of food/nutrient associations related to the disease. That is, in addition to unveiling which nutrient shows significant association with CHD, EWAS also helps us understand which nutrient in connection with which food is responsible for the effect. To demonstrate this principle, we use the food composition table of NHS to extract the contribution of each significant to the total amount of a significant nutrient in the food supply (Fig. 3a). Using a force-directed layout algorithm, we represent this information as a bipartite network, allowing us to explore the significant inter-dependencies among nutrients and food items. In Fig. 3b, negatively associated nutrients and foods are color coded as green and positively associated nutrients and foods are shown in red. Different shapes are used to differentiate between foods (diamonds) and nutrients (circles). The node size is proportional to the absolute value of the estimated effect size. The link thickness is proportional to the amount of each nutrient’s composition in food. Source data are provided in Source Data - Figs. 3-b and 4-b.xlsx.

**Fig. 3** P-value Distribution of Dietary Factors and the Bi-partite Food-Nutrient Network. a P-values are associated with two-sided Wald tests. The Y axis indicates the negative logarithm of the P-value of each exposure. The dotted red horizontal line marks the level of 0.05 P value threshold and the solid red line demonstrates the level of 0.00117 P value threshold associated with the level of 0.05 FDR. Significant foods (diamonds) and nutrients (circles) with a negative association (HR < 1) are highlighted in green and exposures with a positive association (HR > 1) are shown in red. Source data are provided in Source Data - Figure 3-a.xlsx. b Bi-partite food-nutrient network, where negatively associated exposures are shown in green and positively associated exposures are shown in red. Different shapes are used to differentiate between foods (diamonds) and nutrients (circles). The node size is proportional to the absolute value of the estimated effect size. The link thickness is proportional to the amount of each nutrient’s composition in food. Source data are provided in Source Data - Figs. 3-b and 4-b.xlsx.
expected, foods high in negatively associated nutrients are also related to lower CHD risk and foods containing positively associated nutrients are related to higher CHD risk.

In the network shown in Fig. 3b, each node has its own estimated HR. For each nutrient, we compare the estimated HR with the expected geometric hazard ratio \((HR^{N}_{ij})_{ij}\), determined by Eq. (1),

\[
(HR^{N}_{ij})_{ij} = e^{\sum_{k} w_{ik} \beta_{ik}} = \prod_{k=1}^{j} (HR^{F}_{ik})^{w_{ik}},
\]

where \(w_{ik}\) is the normalized weight of the link connecting nutrient
$i$ to food $k$, $\beta^F_1$ is the estimated effect size of food $k$, $HR^F_k$ is the estimated hazard ratio of food $k$, and $HR^N_i \leq (HR^N_i)^*$ (Fig. 4a).

As we show in Fig. 4b, for the majority of nutrients, the actual nutrient HR has the same directionality as the weighted food HR, indicating that EWAS captures both important nutrients and their main drivers in the food system. Whole-grain carbohydrate and milled whole-grain carbohydrate were found to be negatively associated with CHD risk; however, among food items high in these two exposures, we only found doughnuts to relate with higher CHD risk statistically significantly. Consequently, the weighted food HR for these exposures has an opposite directional effect. Yet, the correlation analysis (Supplementary Fig. 3) shows that these two exposures have negative correlations with positively associated foods, such as white bread and processed meats, and positive correlations with negatively associated foods, such as cold breakfast cereal and raw carrots. Moreover, myristic acid, trans-18:2, sodium, and total saturated fat were found to be positively associated with higher CHD risk themselves, but the weighted food HR for them indicates an opposite direction. These nutrients are not only distributed among positively associated foods, but also negatively associated foods. Myristic acid is in both processed meats and yogurt, trans-fatty acid 18:2 is present in salad/oil, vinegar dressing, and doughnuts, sodium is spread among salad/oil and vinegar dressing and processed meats, and total saturated fat is found in yogurt and processed meat. However, the consumption of these exposures is positively correlated with positively associated foods (Supplementary Fig. 3).

The observed disparity indicates that for most of the nutrients, the structure of the food system determines the amount of nutrients in the diet, while for some other nutrients individual choices drive the nutrient amount in the diet. The signal determined by significant foods is a strong driver, but not always sufficient to capture exhaustively the nutrient associations with CHD. Even though some of the nutrients are not well captured by the food approximation (yellow points), with this approach we tend to correctly estimate the sign of the association, while underestimating the effect size. This observation indicates that solely looking at food items, one would underestimate the effect of those nutrients whose consumption is strongly determined by the behavioral aspect and not mainly by their average amount in food. Additionally, this observation can be partially explained by the higher resolution in the calculation of nutrient intake for breakfast cereals, margarine, and types of fats used for cooking and baking, for which we asked separate questions about their...
type and brand that linked with extensive databases for detailed composition values.

Comparison with the literature. Applying the EWAS methodology to NHS data allowed us to identify 53 dietary exposures that show statistically significant association with CHD risk. Next, we discuss our results in the context of the previous literature, offering a direct validation of our findings, also helping detect novel knowledge, and to generate new hypotheses. We found that in most cases, EWAS strengthens the existing knowledge about the effect of diet on CHD, and in some cases, it sheds a light on exposures that have not been thoroughly studied in the literature.

Most of our statistically significant findings were in agreement with the previous literature, strengthening the prior findings and supporting the robustness of the EWAS platform. Among food items, we found that white wine, red wine, and liquor, but not beer, have a negative association with higher CHD risk, aligned with previous findings. Despite recovering an inverse association between alcohol intake and CHD risk (HR 0.88; 95% CI: 0.84–0.91; P value < 5 × 10⁻¹¹), it is worth noting that the overall level of alcohol consumption is not very high in this population study, and such inverse association has not been replicated in other quasi-experimental Mendelian randomization studies.

Moreover, we found salad/oil and vinegar dressing, yogurt, cold breakfast cereal, raw carrots, raisins or grapes, and peanuts have a negative association with higher CHD risk, in line with previous studies. Moreover, we found that total processed meat consumption, hot dogs, apple juice or cider, beverages with sugar, and white bread have positive association with higher CHD risk, as previously shown by other studies.

Many compounds from lipid and fatty acid groups have been studied previously in relation to CHD risk. In line with previous studies, we found that higher consumption of cholesterol, trans-fatty acid 16:1, trans-fatty acid 18:1, trans-fatty acid 18:2, saturated fat, animal monounsaturated fatty acids (MUFA), myristic acid, palmitoleic acid, palmitic acid, and stearic acid are associated with a higher risk of developing CHD (P value < 1 × 10⁻¹³). These dietary factors are mainly distributed among animal-based foods. By contrast, we found that plant MUFA can be protective against CHD development (HR 0.90; 95% CI: 0.85–0.95; P value < 2 × 10⁻⁴), which has also been shown by Zong et al. Plant MUFA are abundant in salad/oil and vinegar dressing and peanuts, food items that we also found to be statistically significantly associated with lower CHD risk (salad/oil and vinegar dressing: HR 0.90; 95% CI: 0.87–0.94; P value < 4 × 10⁻⁷; peanuts: HR 0.93; 95% CI: 0.89–0.97; P value < 5 × 10⁻⁴).

Whole grains are composed of endosperm, germ, and bran, in contrast with milled whole grains in which only endosperm is retained. We found that higher consumption of carbohydrate from whole grains is associated with lower CHD risk, similar to the findings in ref. While the milling process removes several valuable compounds in whole grains, we interestingly detected a similar protective effect for carbohydrate from milled whole grains. In addition, we showed that both natural bran and added bran are negatively associated with CHD risk, consistent with previous studies. We also documented a negative association for cereal fiber with CHD (HR 0.91; 95% CI: 0.87–0.95; P value < 5 × 10⁻⁵), in agreement with ref. One of the food groups that can be rich in natural bran and germ, added bran, and cereal fiber is cold breakfast cereal, which we also found to be negatively related to CHD (HR 0.94; 95% CI: 0.90–0.97; P value < 2 × 10⁻³).

While we rediscovered the negative association of manganese consumption with CHD risk (HR 0.92; 95% CI: 0.88–0.97; P value < 1 × 10⁻³), we also found that higher supplemental selenium is associated with lower CHD risk (HR 0.92; 95% CI: 0.88–0.96; P value < 5 × 10⁻⁴). A specific cardiomyopathy responsive to selenium supplementation has been observed in domestic animals and among Chinese persons with Keshan disease. However, more recent studies found no association between selenium supplementation and primary prevention of cardiovascular disease (CVD), for which reason it is not recommended for CVD prevention. Moreover, our observed positive association of sodium (HR 1.14; 95% CI: 1.08–1.21; P value < 6 × 10⁻⁶) is also consistent with prior studies.

Our findings indicate that higher consumption of dietary folate (HR 0.90; 95% CI: 0.85–0.95; P value < 9 × 10⁻⁶) and folic acid (HR 0.92; 95% CI: 0.88–0.97; P value < 7 × 10⁻⁶) are related to lower CHD risk, again aligned with previous findings. While the beneficial effect of natural vitamin B₆ was previously documented, we interestingly found that increased synthetic vitamin B₆ consumption is related to lower CHD risk (HR 0.94; 95% CI: 0.90–0.97; P value < 2 × 10⁻³). Moreover, our results reveal that beta-tocopherol, total dietary tocopherol intake, alpha-tocotrienol, and beta-tocotrienol are negatively associated with CHD risk (HR < 0.92; P value < 6 × 10⁻⁴). Earlier studies reported similar effects regarding alpha-tocopherol, total tocopherol intake, and alpha-tocotrienol. While the antioxidant and anti-inflammatory effects of some of the vitamin E isomers have been documented, little is known about the effect of beta-tocopherol and beta-tocotrienol on cardiovascular health. These two compounds are mainly found in whole-grain products and nuts. In the liver, beta-tocopherol undergoes omega-hydroxylation, oxidation, and beta-oxidation to generate 13’-hydroxycromanols and 13’-hydroxychromanols, which have potential antioxidant properties. This fact strengthens our findings regarding the protective effect of beta-tocopherol against CHD. Moreover, beta-tocotrienol, in particular, was shown to be inversely related to the risk of type 2 diabetes mellitus, but has not been well-studied regarding its effects on the cardiovascular system. However, among different forms of tocotrienols, beta-tocotrienol has the highest antioxidant activity, in support of the protective effect against CHD that we observed in EWAS.

We found that heme iron (HR 1.08; 95% CI: 1.03–1.14; P value < 2 × 10⁻³) is statistically significantly associated with higher CHD risk. Similar effects were detected in ref. We also foundisorhamnetin and apigenin (HR < 0.92; P value < 2 × 10⁻⁴) to be inversely associated with CHD risk, in line with refs. (Supplementary Table 7). Moreover, in EWAS, we found that higher dietary hydroxyproline intake is associated with higher CHD risk (HR 1.12; 95% CI: 1.06–1.17; P value < 2 × 10⁻⁵). Hydroxyproline is a nonessential amino acid derivative and a major component of the protein collagen mainly found in animal-based food products, such as beef, chicken, and pork. Increased hydroxyproline levels in the urine and/or serum are normally associated with degradation of connective tissue and Marfan syndrome, and were also found to be related to Paget disease. The decrease in various hydroxyproline fractions in aortic tissue of rabbits has been shown to be a risk factor for the progression of atherosclerosis. Nonetheless, serum hydroxyproline is mainly associated with peptides released from the breakdown of collagen, and dietary hydroxyproline intake does not considerably affect serum hydroxyproline levels unless consumed in the form of gelatin. While non-prescription hydroxyproline supplements are available as l-hydroxyproline and N-acetyl-l-hydroxyproline, there seems to be no evidence for the effectiveness of oral hydroxyproline supplements in the prevention or treatment of osteoarthritis, osteoporosis, rheumatoid arthritis, skin ulcers, sports injuries, and wrinkled skin, or in promoting muscle growth or weight loss. The lack of
effectiveness of dietary hydroxyproline is probably a consequence of its failure to be incorporated into collagen: only proline is bioavailable for this purpose. Proline only becomes hydroxylated during a later stage of collagen formation in order to facilitate the strengthening of the collagen helix. Once a collagen helix forms, it does so irreversibly in mammals.\(^8^2\). The ambiguous role of dietary hydroxyproline, along with the positive association with higher CHD risk that we observed in EWAS, emphasizes the need for exploring the metabolic role that this amino acid plays in health.

**Validation in NHS II.** In the second phase of our study, we used NHS II in order to validate statistically significant associations that we found in the original NHS. During 20 years of follow-up in NHS II, 90,861 participants were followed and 604 CHD incidents were documented (Supplementary Methods 3.1). We examined the relationship between 53 exposures found in NHS with CHD risk and deemed an exposure tentatively validated if it had achieved a false discovery rate (FDR) <0.05 significance in NHS and achieved nominal statistical significance in NHS II (\(P\) value < 0.05). Tentatively validated exposures in NHS II had the same directional association with CHD risk as in the original NHS (Table 2). A list of validated and non-validated associations is provided in Supplementary Table 5.

**Discussion**

Our analysis of the dietary determinants of CHD has several limitations based on the nature of the data we used in our analysis. First, the study subjects are only women with a specific occupation (nurses), which restricts the generalizability of the findings to populations comprising males, as well as to more heterogeneous occupational groups and socio-economic backgrounds (Supplementary Figs. 4 and 5). Second, the present study only focused on the effect of dietary factors on CHD risk. Even though diet is an important part of an individual’s environmental exposure, it does not cover the entire exposome\(^8^4\). The environment also includes persistent organic pollutants, plastic-associated chemicals, bacterial and viral infections, air quality, stress, and social network effects\(^8^5\), as well as the endogenous microbiome. Moreover, our diet is not limited to nutrient content; it also carries food additives and other chemicals added during the packaging process, which are absent in the food composition databases, and hence are not included in this.
analysis. Including these chemicals is necessary for a more comprehensive picture of the effect of the diet on health. The only way to overcome these limitations is to include more environmental factors, offering a more comprehensive understanding of the global environment’s effects on health and disease. Moreover, the FFQ used in NHS covers a number of food items that includes the large majority of those consumed by Americans, but the diet of some participants may not be completely represented. Third, while we included the confounding variables that were previously used in other studies on CHD relying on NHS data, our study remains limited to the common confounding variables usually considered when exploring diet–CHD associations. Hence, residual confounding by unmeasured variables cannot be excluded. The obtained results do not unveil causal effects, but, rather, help us generate new hypotheses, which need to be examined in more detail in these and other prospective cohorts and experimental studies. We must also investigate carefully and mechanistically the influence of these dietary factors on human metabolism, exposures that require detailed measurements in terms of dietary bioavailability, hence they can serve as targets for further mechanistic-based analysis.

Our overall goal was to apply GWAS-like analytical approaches to study the dietary determinants of CHD. The methodology allowed us to explore both food items and nutrients, offering a more comprehensive picture of the effect of diet on CHD and helping us visualize the obtained relationships using network tools. Our study not only reproduced the prior knowledge in the diet–CHD domains, but also led to novel associations. While some of the previous EWAS studies failed to achieve adequate statistical power in association detection, our positive results suggest that these failures were often related to the cohort size and the absence of repeated longitudinal dietary assessments. Indeed, our use of a large longitudinal dataset with a long follow-up period and a sufficient number of subjects helped us achieve sufficient statistical power to detect even relatively small effect sizes. However, enabling a wide-association study to investigate environmental factors requires careful consideration in designing cohort studies, and detailed, comprehensive exposure assessment methods to ensure that the effect of the environment is fully captured. While selectively testing and reporting one or a few associations has been argued to be a source of biased results and false positives, there is clearly a role for testing specific etiologic hypotheses as this allows greater statistical power and a more detailed examination of an exposure–outcome relationship. An environment-wide association study is a complementary approach that allows us to rank the associations and report transparently both significant and non-significant associations. It also allowed us to generate new hypotheses that can be further investigated in single-association studies and mechanism-based studies. It is worth mentioning that until recently NIH was unlikely to fund research proposals without well-developed hypotheses, restricting the possibility of conducting wide-association studies and this analysis was only possible because of the accrual of over 30 years of follow-up in this large cohort.

In the present study, we explored the effect of only 374 dietary exposures. Yet, when it comes to the chemical composition of the food we consume, these nutritional components represent only a tiny fraction of the thousands of distinct definable biochemicals that have been identified in food. While many of these chemicals have well-documented or potential implications for health, they remain largely unquantified in any systematic fashion across different individual foods. Their invisibility to experimental, clinical, epidemiological, and demographic studies—turning them into the virtual dark matter of nutrition research—represents a roadblock toward a better, more consistent, more reproducible understanding of how diet affects health. The high-resolution diet description space, the conventional single-association approach is even more impractical and lacks scalability. The EWAS methodology, however, would be able to test higher order of magnitude of dietary compounds, to identify significant associations with a disease of interest or with a prescription for health.

Methods

Knowledge graph. To create the knowledge graph, we firstly identified in PubMed publications that have NHS or Health Professionals Follow-up Study in the title or abstract, along with papers co-authored by the main PIs of the NHS. We manually filtered papers that studied the association between dietary exposures and cardiovascular complications, such as coronary heart disease, stroke, and hypertension. Since not all papers are indexed on PubMed, we searched the web using the same database. The remaining 2688 filtered papers that had both exposure and outcome in the titles and abstracts of the obtained papers and extracted the exposure–phenotype relations, the associations found, the effect size, and other related information (Table 3). More than one association might be studied in a paper. Overall, we found 292 studied associations documented in 91 papers, altogether 124 negative and 45 positive associations were documented in relation to cardiovascular complications. In the remaining cases, there was no significant association between an exposure and a phenotype of study. The obtained data are shown in a knowledge graph (Fig. 1) where each link represents an association. The space of studied exposures is rather heterogeneous and is often driven by the researcher’s interests and experience. For example, the two nutrients with respect to CHD mortality, 20600 or >3550 kcal/day, or left >70 questions in the FFQ unanswering. Participants received one questionnaire every 2 years to report their dietary data and one questionnaire every 4 years to document their dietary intake. At any point within the follow-up period, if a participant reported development of non-fatal myocardial infarction or fatal CHD10,101,102, she will be classified as a case, with no further update of her dietary records. If she developed other diseases, such as diabetes mellitus or cancer, she would still be classified as a non-case, with no further update of her dietary records (Supplementary Methods 3.1 and Supplementary Fig. 1). These exclusion criteria were chosen by virtue of minimizing reverse causation bias from death. As we did not focus on the impact of measurement errors and missing data, we included 62,811 subjects in the analysis, representing 2774 cases (4%) and 60,037 controls (96%).

Ascertainment of diet. NHS uses an internally designed FFQ, with documented reproducibility and validity. The FFQ has been regularly updated to adapt to changes in the food market and to capture additional food items. For each food item, the FFQ specified a commonly used unit or portion size, asking each subject how often, on average, she had consumed that quantity during the past year. Nine responses were possible, ranging from “almost never” to “six or more times per day.” We converted the frequency responses to the number of servings per day for each food item. We calculated daily intake of nutrients by multiplying the frequency of consumption of each item by its nutrient content and summing the nutrient contributions of all foods on the basis of Harvard University Food Composition Database derived from US Department of Agriculture sources and other resources, including published reports, data from manufacturers, and in-house analyses of fatty acid composition. We looked into several food items more closely. For example, we used an algorithm designed by Jacobs et al. to classify breakfast cereals into wholegrain and refined grain. We also collected detailed information on the type of fat or oil used in food preparation and brand or type of margarines to calculate the fatty acid consumption.

Ascertainment of CHD. We ascertained incident cases of CHD (non-fatal MI or fatal CHD) that occurred after the return of the 1986 questionnaire but before June 30, 1990. Physicians, unaware of the self-reported risk factor status, systematically reviewed the medical records of those who reported having an MI on each biennial questionnaire. MI was classified as confirmed if the World Health Organization criteria, that is, symptoms, electrocardiographic changes, or elevated cardiac enzyme concentrations, were met. Fatal CHD was confirmed by either hospital pathologist or through an autopsy. If it was listed as the cause of death on the death certificate, if it was listed as an underlying and most plausible cause of death, or if evidence of previous CHD was available. Deaths were identified from state vital
Table 3 Knowledge graph data preparation.

| Column | Cox | 2018 | Carbohydrate quality and quantity and the risk of coronary heart disease among US
| Phenotype | Exposure | Year | Title | Abstract | Gender | Association | Model | Effect size |
|----------|---------|------|-------|---------|--------|-------------|-------|------------|
| Total fiber | Cereal fiber ratio | 2018 | Carbohydrate quality and quantity and the risk of coronary heart disease among US
| Gender | Total fiber | Cox | 0.8 | Positive | Both | No association | CHD | Cox |
| Total fiber | Cereal fiber ratio | 2018 | Carbohydrate quality and quantity and the risk of coronary heart disease among US
| Gender | Total fiber | Cox | 1.2 | Positive | Both | No association | CHD | Cox |

We aimed to assess the relation between various measures of carbohydrate quality and the risk of coronary heart disease (CHD) among US

We adjusted for potential risk factors and confounders, including age (Supplementary Fig. 2), BMI, physical activity, and total caloric intake as con

The authors declare that all data supporting the findings of this study are available upon request to Nurses’ Health Study (NHS) and when the request for data access is approved. Written informed consent was obtained from participants to release medical records documenting the incidence of coronary heart disease.

Data availability

The programming materials are available on GitHub and Zenodo. R version 3.4.0 was used for data analysis and Python version 2.7.16 and MATLAB 2019a were used for data visualizations.

Received: 21 June 2019; Accepted: 29 October 2020; Published online: 27 November 2020
References

1. Siscovick, D. S. et al. Dietary intake and cell membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. JAMA 274, 1363–1367 (1995).
2. Estruch, R. et al. Primary prevention of cardiovascular disease with a Mediterranean diet. N. Engl. J. Med. 368, 1279–1290 (2013).
3. Li, S. et al. Better diet quality and decreased mortality among myocardial infarction survivors. JAMA Intern. Med. 173, 1808–1819 (2013).
4. Mente, A., de Koning, L., Shannon, H. S. & Anand, S. S. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. Archiv. Intern. Med. 169, 659–669 (2009).
5. Division for Heart Disease and Stroke Prevention. Heart Disease and Stroke Maps www.cdc.gov/heartdisease/maps (2016).
6. Robertson, T. L. et al. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: coronary heart disease risk factors in Japan and Hawaii. Am. J. Cardiol. 39, 244–249 (1977).
7. Capuano, E., Oliviero, T. & van Boekel, M. A. Modeling food matrix effects on chemical reactivity: challenges and perspectives. Crit. Rev. Food Sci. Nutr. 58, 1–15 (2018).
8. Kolonel, L. N., Yoshizawa, C. N. & Hankin, J. H. Diet and prostatic cancer: a case-control study in Hawaii. Am. J. Epidemiol. 127, 999–1012 (1988).
9. Marchand, L. L., Hankin, J. H., Kolonel, L. N. & Wilkens, L. R. Vegetable and fruit consumption in relation to prostate cancer risk in Hawaii: a reevaluation of the effect of dietary beta-carotene. Am. J. Epidemiol. 133, 215–219 (1991).
10. Mente, A., de Koning, L., Shannon, H. S. & Anand, S. S. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. Archiv. Intern. Med. 169, 659–669 (2009).
11. Tapsell, L. C., Neale, E. P., Satija, A. & Hu, F. B. Foods, nutrients, and dietary patterns: interconnections and implications for dietary guidelines. Adv. Nutr. 7, 445–456 (2016).
12. Hu, F. B. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr. Opin. Lipidol. 13, 3–9 (2002).
13. Shimazu, T. et al. Dietary patterns and cardiovascular disease mortality in a Japanese prospective cohort study. Int. J. Epidemiol. 5, 255–269 (2007).
14. Patel, C. J., Bhattacharya, J. & Butte, A. J. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS ONE 5, e10746 (2010).
15. Balan, M. I. & Daviglus, M. L. Dietary and plasma magnesium and risk of coronary heart disease among women. N. Engl. J. Med. 348, 109–118 (2003).
16. Estruch, R. et al. Moderate consumption of red wine, but not gin, decreases erythrocyte superoxide dismutase activity: a randomised cross-over trial. Nutr. Metab. Cardiovasc. Dis. 21, 46–53 (2011).
17. Millerud, I. Y. et al. Conventional and genetic evidence on alcohol and vascular disease aetiology; a prospective study of 500 000 men and women in china. Lancet 393, 1831–1842 (2019).
18. Holmes, M. V. et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. BMJ 349, g4164 (2014).
19. Hu, F. B. et al. Dietary intake of α-linolenic acid and risk of fatal ischemic heart disease among women. Am. J. Clin. Nutr. 69, 890–897 (1999).
20. Rice, B. H. Dairy and cardiovascular disease: a review of recent observational research. Curr. Nutr. Rep. 3, 130–138 (2014).
21. Djoussé, L. & Gagnon, J. M. Breakfast cereals and risk of heart failure in the physicians’ health study I. Arch. Intern. Med. 167, 2080–2085 (2007).
22. Gagnon, J. M. et al. A prospective study of consumption of carotenoids in fruits and vegetables and decreased cardiovascular mortality in the elderly. Am. J. Epidemiol. 172, 591–598 (2010).
23. Puglisi, M. J. et al. Raisins and additional walking have distinct effects on plasma lipids and inflammatory cytokines. Lipids Health Dis. 7, 14 (2008).
24. Hu, F. B. et al. Frequent nut consumption and risk of coronary heart disease in women: prospective cohort study. BMJ 317, 1341–1345 (1998).
25. Micha, R., Machas, G. & Mozaffarian, D. Unprocessed red and processed meats and risk of coronary artery disease and type 2 diabetes—an updated review of the evidence. Curr. Atheroscler. Rep. 14, 515–524 (2012).
26. Pase, M. P., Grima, N., Cockerell, R. & Pipingas, A. habitual intake of fruit juice predicts central blood pressure. Appetite 84, 68–72 (2015).
27. Fung, T. T. et al. Sweetened beverage consumption and risk of coronary heart disease in women. Am. J. Clin. Nutr. 89, 1037–1042 (2009).
28. Sieri, S. et al. Dietary glyceric load and index and risk of coronary heart disease in a large Italian cohort: the epic study. Arch. Intern. Med. 170, 640–647 (2010).
29. Connor, W. E. & Connor, S. L. Dietary cholesterol and coronary heart disease. Circ. Atherosclerosis Rep. 4, 425–422 (2002).
30. Mozaffarian, D., Aro, A. & Willett, W. C. Health effects of trans-fatty acids: experimental and observational evidence. Eur. J. Clin. Nutr. 65, S5 (2009).
31. Sun, Q. et al. A prospective study of trans fatty acids in erythrocytes and risk of coronary heart disease. Circulation 115, 1858–1865 (2007).
32. Zong, G. et al. Intake of individual saturated fatty acids and risk of coronary heart disease in us men and women: two prospective longitudinal cohort studies. BMJ 355, g5796 (2016).
33. Zong, G. et al. Monounsaturated fats from plant and animal sources in relation to risk of coronary heart disease among us men and women. Am. J. Clin. Nutr. 107, 445–453 (2018).
34. Djoussé, L., Weir, N. L., Hanson, N. Q., Tsai, M. Y. & Gagnon, J. M. Plasma phospholipid concentration of cis-palmitoleic acid and risk of heart failureclinical perspective. Circ. Heart Fail. 5, 703–709 (2012).
35. Hu, F. B., Manson, J. E. & Willett, W. C. Types of dietary fat and risk of coronary heart disease: a critical review. J. Am. Coll. Nutr. 20, 5–19 (2001).
36. Li, Y. et al. Saturated fats compared with unsaturated fats and sources of carbohydrates in relation to risk of coronary heart disease: a prospective cohort study. J. Am. Coll. Cardiol. 66, 1538–1548 (2015).
37. Wu, H. et al. Association between dietary whole grain intake and risk of mortality: two large prospective studies in US men and women. JAMA Intern. Med. 175, 373–384 (2015).
38. Jensen, M. K. et al. Intakes of whole grains, bran, and germ and the risk of coronary heart disease in men. Am. J. Clin. Nutr. 80, 1492–1499 (2004).
39. ALESSA, H. B. et al. Carbohydrate quality and quantity and risk of coronary heart disease among women and men. Am. J. Clin. Nutr. 107, 257–267 (2018).
40. Rosique-Esteban, N., Guasch-Ferré, M., Hernández-Alonso, P. & Salas-Salvado, J. Dietary magnesium and cardiovascular disease: a review with emphasis in epidemiological studies. Nutrients 10, 168 (2018).
Acknowledgements
We were supported by grants from NIH (grants P01 HL132825, UM1 CA186107, P01 CA87969, R01 CA49449, R01 HL034594, R01 HL088521, UM1 CA176726, R01 CA67262, U54 HL119145, U01 HG007690, and P50 GM107618) and AHA (grants 151708, 414110-68953, and D700382). A.-L.B. was supported by NIH 1P01HL132825, Rockefeller Foundation 2109 FOD 026, and the European Union’s Horizon 2020 research and innovation programme under grant agreement No 810115 - DYNASNET.

Author contributions
S.M. performed data query and integration, statistical modeling, network analysis, and programming and contributed to writing the manuscript. G.M. contributed to network analysis, statistical modeling, programming, and writing the manuscript. Y.L. contributed to data query and programming. J.L. contributed to interpreting the results and writing the manuscript. W.C.W. contributed to data collection, analyzing the results, and writing the manuscript. A.-L.B. contributed to the conceptual design of the study and writing the manuscript.

Competing interests
A.-L.B. is founder of Nomix and Foodome, and J.L. and A.-L.B. are founders of Scipher Medicine, companies that explore the use of network-based tools in health. The remaining authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19888-2.

Correspondence and requests for materials should be addressed to A.-L.B.

Peer review information Nature Communications thanks Ramon Estruch, Paolo Vineis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020