nature food

Article

https://doi.org/10.1038/s43016-025-01147-6

Dietary species richness provides a comparable marker for better nutrition and health across contexts

Received: 21 January 2024

Accepted: 21 February 2025

Published online: 24 March 2025

Check for updates

Giles T. Hanley-Cook[1,28] ✉, Jill Deygers[1,28], Aisling J. Daly[2], Jeroen Berden[1,3], Roseline Remans[4], Celine Termote[5], Daniel B. Ibsen[6,7,8,9], Julia Baudry[10], Patrick Van Damme[11], Emmanuelle Kesse-Guyot[10], Paolo Vineis[12], Matthias B. Schulze[13,14], Ky The Hoang[15], Mélanie Deschasaux-Tanguy[10], Alicia Heath[16], Christina C. Dahm[6], Yvonne T. van der Schouw[17], Guri Skeie[3,18], Marcela Guevara[19,20,21], Lorenzo Milani[22], Daniela Penafiel[23], Jessica E. Raneri[1], Francis Odhiambo Oduor[5,24], Danny Hunter[25], Disna Ratnasekera[26], Kris A. Murray[27], Mathilde Touvier[10], Inge Huybrechts[3] & Carl Lachat[1] ✉

Ecological diversity indices such as Hill numbers have been developed to estimate effective species numbers, yet the ability of Hill numbers to compare food biodiversity across contexts is unclear. Here we computed the between- and within-country variability of similarity-insensitive Hill numbers using dietary intake collected from prospective cohorts in nine European countries and cross-sectional studies in five low- and middle-income countries. We also assessed the relationships between more biodiverse diets, mortality rates and micronutrient adequacy. Only $Hill_0$, better known as dietary species richness (DSR), showed strong heterogeneity between countries and individuals within countries. Higher DSR was most strongly associated with lower mortality rates in Europe as compared to $Hill_1$, $Hill_2$ and $Hill_\infty$, whereas relationships with micronutrient adequacy were comparable across Hill numbers in the global south. DSR can be used to assess progress towards more biodiverse diets, while also serving as a marker for the deleterious nutrition and health impacts associated with non-diverse diets.

Industrial agriculture production systems are based on a narrow range of crop and animal species, as reflected by fewer than 200 plant and 40 animal species making substantial contributions to global food production[1,2], and have been associated with the homogenization of diets[3], high rates of malnutrition, such as micronutrient deficiencies and obesity[4,5] and accelerating biodiversity loss globally[6]. Evidence on the human and planetary health co-benefits associated with more diversified agrifood systems[7,8], food supplies[9] and subsequent diets is unambiguous[10–12].

In recognition that food, not nutrients, is the functional unit in nutrition[13], dietary diversity has traditionally been quantified by counting the number of distinct food groups consumed, which include among others pulses, vegetables and seafood[14]. However, food group diversity indices fail to reflect the inherent variability in chemical components, such as nutrients but also bioactive non-nutrients[15], and ecological benefits, such as resource partitioning and soil biodiversity[16], delivered by edible species within food groups (for example, carrot, pumpkin) or the relative distribution, also referred to as moderation, of species across a diet[17].

Therefore, the concept of food biodiversity, defined as the diversity of plants, animals and other organisms between and within food groups consumed as food and drink, has drawn attention as a more

A full list of affiliations appears at the end of the paper. ✉e-mail: Giles.HanleyCook@UGent.be; Carl.Lachat@UGent.be

https://doi.org/10.1038/s43016-025-01147-6

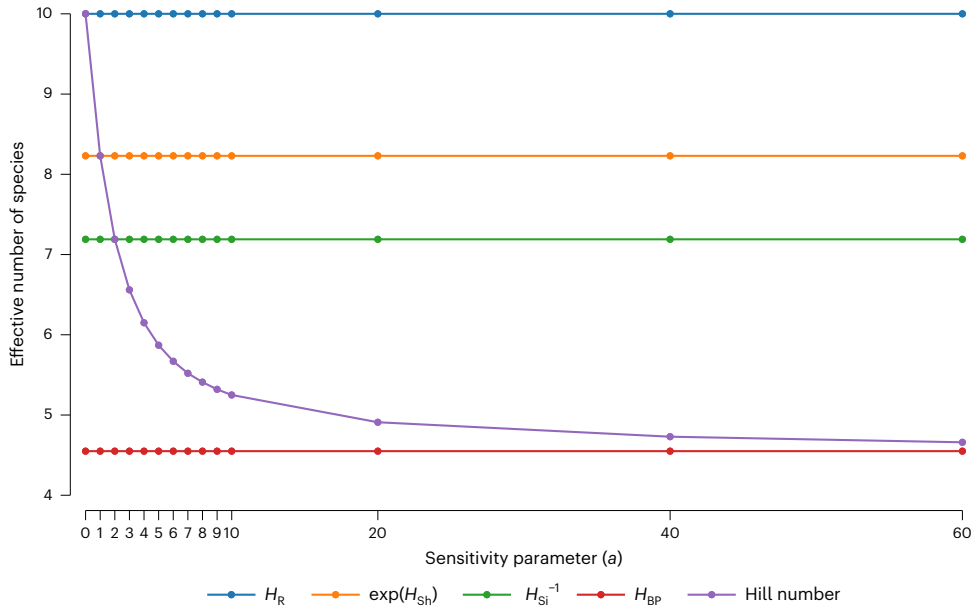

Fig. 1 | Comparison of similarity-insensitive diversity indices for a hypothetical diet. The assumed abundance vector for the ten species is **p**= (0.22, 0.19, 0.14, 0.11, 0.09, 0.07, 0.07, 0.06, 0.03, 0.02). Hill numbers result in diversity profiles, whereas univariate indices are not functions of a. Values of $a \geq 3$ are shown to illustrate the convergence towards H_{BP}.

holistic entry point to help define, and subsequently achieve, sustainable healthy diets worldwide. However, assessments of (bio)diversity—both within and across scientific fields—have used a plethora of diversity indices[18], which have distinct data requirements, functional units[19], mathematical properties[20] and thus nutritional and ecological interpretations[21].

To illustrate, in nutrition, dietary richness (H_R) is conventionally expressed as the number of unique species consumed, regardless of quantities[22,23]; the Shannon index (H_{Sh}) is a measure of the entropy of a diet or, in other words, the number of unique species and the relative quantities of these species consumed[24]; the Simpson index (H_{Si}) is the probability that two foods taken at random (with replacement) from a diet represent the same species[25]; and the Berger–Parker index (H_{BP}) is the reciprocal of the proportion of the most common species consumed[3]. The key drawbacks of these classical diversity indices are that direct comparisons between diets using distinct indices, with different functional units, or comparisons between diets using the same index, due to the nonlinearity of one-unit increments in H_{Sh} or H_{Si}, are often unfounded through the lens of nutrition[21].

Previously, Hanley-Cook et al.[21] postulated that Hill numbers, also known as effective species indices[26,27], might help circumvent the limitations and widespread misinterpretation of well-established diversity indices used in nutritional epidemiology. To clarify, converting a diet to its effective number of species entails finding an equivalent diet (that is, same value of the diversity index as the initial diet in question) in which each species is consumed in equal quantities[20]. Hence, Hill numbers could facilitate a unification of food biodiversity indices through their analogous functional unit, known as the effective number of species. However, at present, there is a lack of empirical evidence confirming whether the use of Hill numbers—which place a varying degree of weight on less or more abundant species in the diet—would allow for rational comparisons between different study contexts, dietary assessment methods (for example, food frequency questionnaires vs 24-hour recalls) and/or food biodiversity indices[21].

Therefore, the primary objective of this study is to compute and compare the between- and within-country variability in similarity-insensitive Hill numbers using quantitative dietary intake data collected among ~450,000 adults in Europe and ~2,200 females in low- and middle-income countries (LMICs) in Africa, Asia and South America. In addition, to further assess the relevance of food biodiversity for public health nutrition (for example, sustainable food-based dietary guidelines)[28], we examined the magnitude of the relationships between Hill numbers and rates of all-cause mortality in nine European countries and micronutrient adequacy in five LMICs.

Results

The mathematical relationships between similarity-insensitive Hill numbers for a hypothetical diet are illustrated in Fig. 1. Further details are provided in the Methods.

Hill number distributions

The number of effective species in diets varied substantially across Hill numbers (Fig. 2). In both the European Prospective Investigation into Cancer and Nutrition (EPIC) and LMIC studies, $Hill_0$ (that is, H_R) exhibited the largest heterogeneity between countries and among participants within countries, as reflected by the more differentiated and wider distributions, while $Hill_\infty$ showed the least variability, in other words indiscriminate and narrow distributions (Fig. 2).

Our comparative analyses, using Hill numbers based on relative energy intakes (kcal per day), confirmed these findings (Supplementary Fig. 1). Furthermore, higher values of Hill numbers were not associated with consuming greater amounts of food and drink (g per day) or dietary energy (kcal per day) (Supplementary Fig. 2).

Rank correlations across Hill numbers

In both the EPIC and LMIC studies, $Hill_1$, $Hill_2$, and $Hill_\infty$ showed strong to very strong Spearman's rank correlations (that is, ρ: 0.79–0.97). In contrast, in EPIC, $Hill_0$ was only weakly to very weakly correlated with the other Hill numbers (that is, ρ: 0.04–0.22), while in the LMIC study, moderate to strong correlations were observed (that is, ρ: 0.51–0.75) (Fig. 3).

Similar findings were observed when correlating Hill numbers based on relative energy intakes (kcal per day), although in the EPIC study, energy-based $Hill_0$ was more strongly correlated with the other Hill numbers (that is, ρ: 0.15–0.39) (Supplementary Fig. 3). Furthermore, in the LMIC study, both weight- and energy-based Hill numbers

https://doi.org/10.1038/s43016-025-01147-6

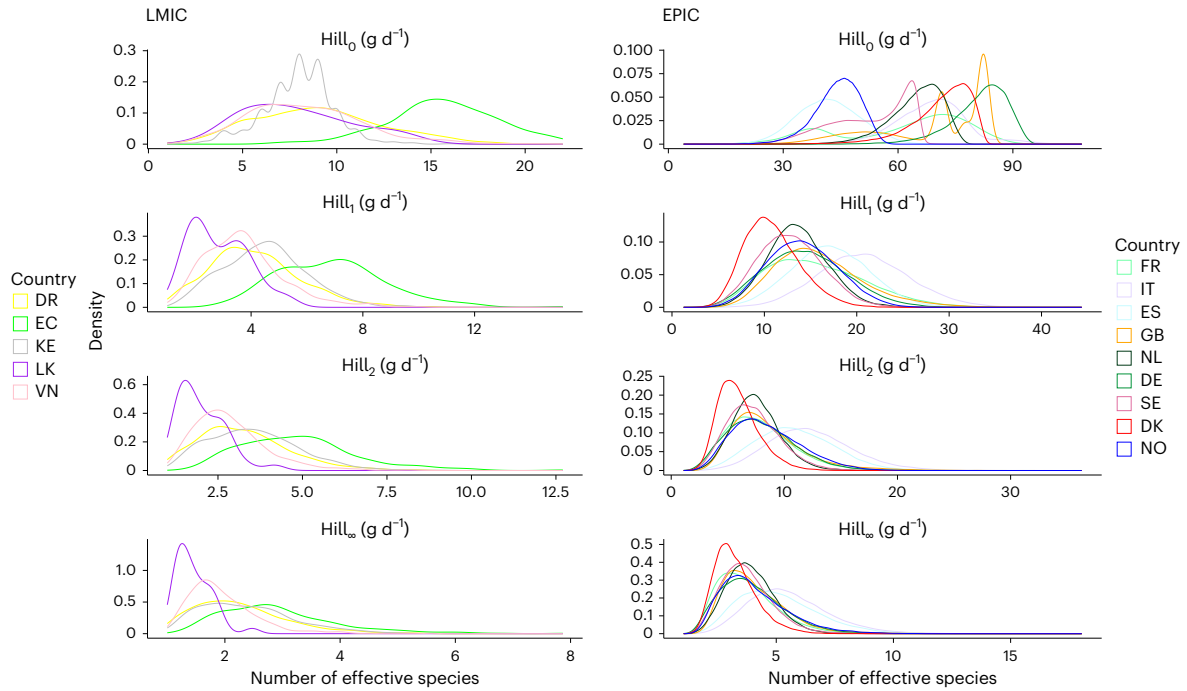

Fig. 2 | Distribution of Hill numbers in the EPIC and LMIC studies, by country. DE, Germany (n = 49,352); DK, Denmark (n = 55,014); DR, Democratic Republic of Congo (n = 461); EC, Ecuador (n = 251); EPIC, European Prospective Investigation into Cancer and Nutrition; ES, Spain (n = 39,990); FR, France (n = 67,920); GB, The United Kingdom of Great Britain and Northern Ireland (n = 75,372); IT, Italy (n = 44,547); KE, Kenya (n = 770); LK, Sri Lanka (n = 36); LMIC, low- and middle-income country; NL, The Netherlands (n = 36,538); NO, Norway (n = 33,967); SE, Sweden (n = 48,690); VN, Viet Nam (n = 631). Data availability section provides access details to the source data.

showed strong to very strong rank correlations (that is, ρ: 0.64–0.78), whereas in the EPIC study, only moderate to very low correlations were observed among Hill$_1$, Hill$_2$ and Hill$_\infty$ (that is, ρ: 0.12–0.39). Naturally, weight- and energy-based Hill$_0$ were perfectly correlated, as their computations are independent of relative quantities of species consumed (Supplementary Fig. 4).

Associations between Hill numbers and all-cause mortality

For each Hill number, higher quintiles (Qs) were inversely associated with all-cause mortality rates over the ~20 years of follow-up (Fig. 4 and Supplementary Table 1). The strongest and most uniform relative reduction in death rates (that is, multivariable-adjusted hazard ratios (HRs), using the first Q as the reference) was observed across Qs of Hill$_0$. However, the highest absolute reduction in mortality rates, which is an estimate unadjusted for potential confounding variables, was observed across Qs of Hill$_1$ (that is, 57 fewer deaths per 10,000 person years, when comparing Q$_5$ against Q$_1$).

Our comparative analyses, using Hill numbers based on relative energy intakes (kcal per day), confirmed our main findings on relative reductions in death rates (Supplementary Fig. 5 and Supplementary Table 2). However, the absolute reduction in mortality rates were weaker for energy-based Hill$_1$ and Hill$_2$, whereas the death rate was in fact higher in Q$_5$ than Q$_1$ for energy-based Hill$_\infty$, but these estimated are prone to a potentially larger magnitude of confounding.

In addition, when excluding other dietary risk factors from our models, thus assuming them to be potential mediators, rather than confounders, higher values of Hill numbers were more strongly associated with reductions in death rates (Supplementary Tables 3 and 4). Lastly, including the percent of kcal per day intakes from ultra-processed foods, widely referred to as Nova group 4, either in addition to or as a substitute for other dietary risk factors, did not significantly change the estimated HRs (Supplementary Tables 5 and 6).

Association between Hill numbers and micronutrient adequacy

For each Hill number, higher quartiles were positively associated with greater mean adequacy ratios (MARs) (Fig. 5 and Supplementary Table 7). The strongest and most uniform increments were observed across quartiles of Hill$_1$ (that is, 95% confidence interval (CI): 12.2 to 15.3 percentage point greater MAR, when comparing the highest quarter against the lowest quarter). However, higher quartiles of Hill$_0$, Hill$_2$ and Hill$_\infty$ were also strongly and statistically equivalently associated with higher micronutrient adequacy.

Our comparative analyses, using Hill numbers based on relative energy intakes (kcal per day), confirmed our main findings (Supplementary Fig. 6 and Supplementary Table 8). Nevertheless, energy-based Hill$_2$ performed less adequately but comparably to Hill$_\infty$.

Discussion

In this study, we report that Hill numbers, with the notable exception of Hill$_0$, poorly discriminated countries and individuals within countries, even within studies with very large number of participants such as EPIC. These findings raise doubt over the descriptive utility, and potential sensitivity to change, of weight- and energy-based Hill$_1$, Hill$_2$ and Hill$_\infty$ for smaller-scale studies assessing individual-level quantitative dietary intake and support the use of dietary species richness (that is, Hill$_0$ or H_R). This recommendation is further strengthened by the results of our comparative epidemiological analyses. In other words, regardless of which Hill number was used as the dietary exposure in the EPIC or LMIC studies, the same overarching conclusion was drawn that higher food biodiversity is associated with lower rates of all-cause mortality and higher micronutrient adequacy, respectively. In fact, Hill$_0$ was most strongly associated with lower all-cause mortality in Europe, whereas associations with more adequate micronutrient intakes in LMICs were comparable across Hill numbers. These findings indicate that less abundant species might have important nutrition and health benefits or act

https://doi.org/10.1038/s43016-025-01147-6

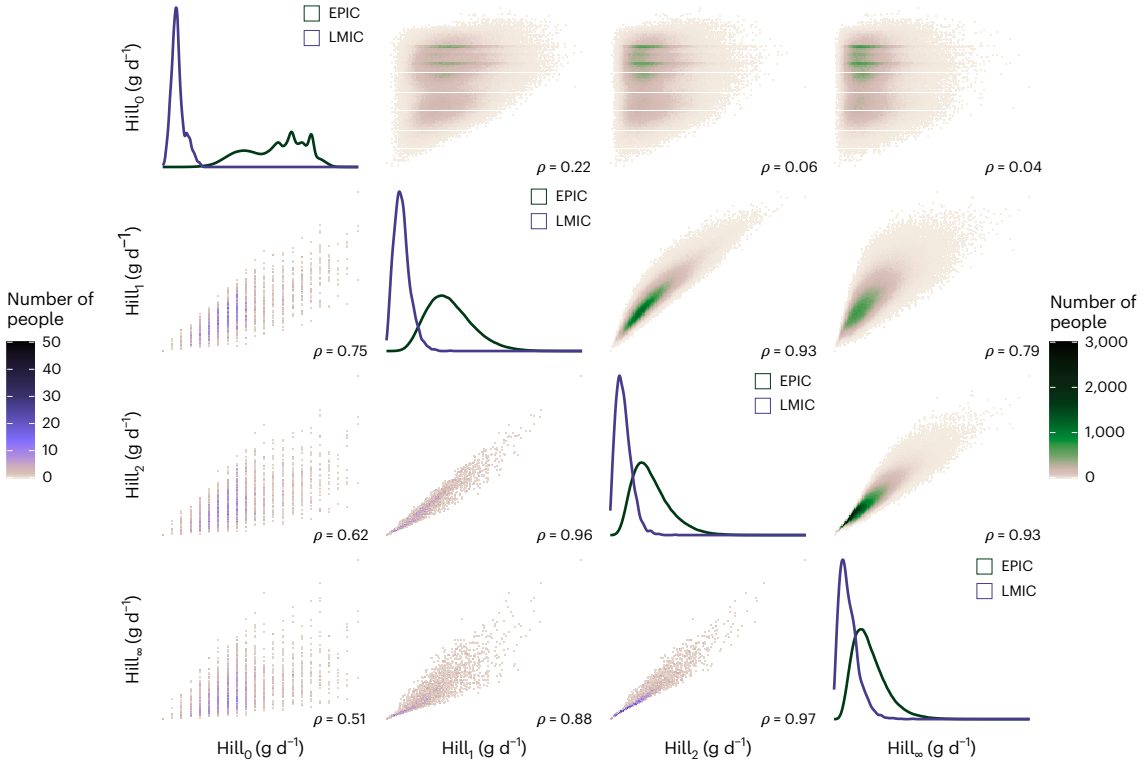

Fig. 3 | Heatscatters and correlations among Hill numbers in the EPIC and LMIC studies. EPIC, European Prospective Investigation into Cancer and Nutrition (n = 451,390); LMIC, low- and middle-income country (n = 2,186); ρ, Spearman's rank correlation coefficient. Data availability section provides access details to the source data.

as proxies for unmeasured healthy behaviours. Furthermore, although higher values of Hill$_1$, Hill$_2$ and Hill$_\infty$ were linked to lower absolute death rates than Hill$_0$, these estimates are unadjusted for potential confounders and therefore lacking meaningful causal interpretations. Therefore, for large-scale, robust and timely surveillance of food biodiversity across contexts and time, non-quantitative assessments of Hill$_0$, which is a simple count of the unique species consumed by individuals, are probably sufficiently sensitive to gauge progress on food biodiversity and foreshadow adverse nutritional impacts.

Nonetheless, for research purposes and triangulation, a suite of complementary Hill numbers is appropriate, including Hill$_1$, Hill$_2$ and Hill$_\infty$, which would require the resource-intensive collection of abundance data to determine the percent of g per day or kcal per day intakes per species. Future studies might assess whether using the assumption of uniform dietary energy intake for each study participant, for example, 2,500 kcal per day, in combination with a semi-quantitative screener of species usually contributing large shares of food energy, such as common wheat, potato, maize, cow, pig, is a valid and scalable method for energy-based Hill$_\infty$. This methods would require only the relative abundance of the most common species consumed to compute Hill$_\infty$, which was strongly correlated with Hill$_1$ and Hill$_2$. For relative comparisons of food biodiversity in smaller dietary intake studies, standardization tools such as coverage-based rarefaction and extrapolation of datasets are warranted to obtain robust estimates of Hill numbers[29].

The potential mechanisms driving the associations between food biodiversity, such as Hill numbers, and nutrition and health outcomes have been described previously[21,22]. In brief, the first process is coined as the 'sampling effect,' which assumes that simply by chance more diverse diets lead to the inclusion of highly nutritious (for example, micronutrient rich) and health protective species (for example, containing a plethora of bioactive non-nutrients). To illustrate, it has been estimated that there are between 100,000 and 1,000,000 plant metabolites alone[30]. Notably, participatory global initiatives such as the Periodic Table of Food Initiative aim to expand food composition and function data[31] to include both nutrients and the unmapped complexity of other bioactive components, such as phytochemicals[15].

The second mechanism is known as the 'complementarity effect', which postulates that more biodiverse diets contain species with complementary nutritional profiles to cover all requirements[32,33] and that the chemical or physical interactions between food species and their components result in a function greater than expected by chance, as illustrated by vitamin C increasing the absorption of non-haem iron.

The third process is referred to as 'minimizing trade-offs', which occurs by more diverse diets mitigating excessive intakes of a limited number of food groups or edible species. For example, diets composed of large quantities of cruciferous vegetables may lead to unsafe exposure to dietary goitrogens.

The fourth process is the 'microbiome effect', which hypothesizes a mediating role for microbial communities in nutrition and health disparities. To clarify, higher dietary diversity has been associated with more diverse gut microbial communities[34], which have in turn been linked with a number of metabolic, nutritional and health benefits[35,36].

In this study, we observed imperfect concordance between weight- and energy-based Hill$_1$, Hill$_2$ and Hill$_\infty$ in relation to one another and with nutrition and health outcomes. In other words, species that contribute relatively large quantities to a diet (that is, percent of g per day), such as fruit and vegetable species, do not necessarily contribute substantially towards energy intakes (that is, percent of kcal per day) or vice versa, as is the case for animal species[22]. Notably, weight-based Hill$_1$, Hill$_2$ and Hill$_\infty$ showed stronger relationships with micronutrient adequacy than their energy-based counterparts. This observation is probably attributable to the highly variable food composition (that is, micronutrient contents and densities) of distinct species, in particular those

https://doi.org/10.1038/s43016-025-01147-6

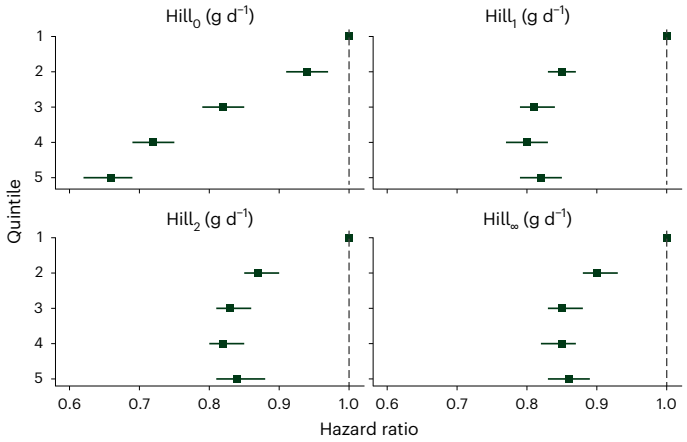

Fig. 4 | Associations between higher quintiles of Hill numbers and all-cause mortality rate in the EPIC study. Multi-adjusted models were stratified for centre, age at recruitment (1-year intervals, timescale) and sex and adjusted for baseline alcohol intake (g per day), physical activity (Cambridge index: active; moderately active; moderately inactive; inactive; missing), marital status (single, divorced, separated or widowed; married or living together; unknown), smoking status and intensity of smoking (current, 1 to 15 cigarettes per day; current, 16 to 25 cigarettes per day; current, 26+ cigarettes per day; current, pipe/cigar/occasional; current/former, missing; former, quit 11 to 20 years; former, quit >20 years; former, quit ≤10 years; never; unknown), educational level (longer education (including university degree, technical or professional school); secondary school; primary school completed; not specified), baseline energy intake (kcal per day), baseline fibre intake (g per day), baseline red and processed meat consumption (g per day) and an 18-point Mediterranean diet score. Error bars represent the 95% confidence intervals around the point estimate. EPIC, European Prospective Investigation into Cancer and Nutrition (n = 451,390). Data availability section provides access details to the source data.

habitually contributing small relative quantities to the total weight but high relative shared of the total energy of a diet, such as cow, salmon and peanuts for iron, zinc and folate intakes, respectively. In other words, higher values of weight-based Hill$_1$, Hill$_2$ and Hill$_\infty$ probably act as a proxy for greater intakes of micronutrient-dense species.

Nonetheless, as is widely acknowledged, micronutrient intakes in isolation fail to determine the healthfulness of whole dietary patterns[37]. This might clarify why the strong relationships between weight-based Hill numbers and MAR in the LMIC study did not carry over to meaningfully larger reductions in death rates in the EPIC study, when compared to energy-based Hill numbers. Moreover, weak to moderate correlations between weight- and energy-based Hill numbers observed in the EPIC study hamper their relevance for targeting individuals, as depending on the abundance unit used to determine relative consumption quantities (that is, g per day or kcal per day), the same person could be classified as being either a low or high consumer of food biodiversity.

Therefore, Hill numbers can be regarded as a critical step forward towards unifying similarity-insensitive diversity indices, through effective number of species as the functional unit, but not a panacea for the quantification of food biodiversity. Indeed, we must acknowledge that Hill numbers using relative abundances in their computation (for example, percent g per day) remain challenging to interpret through the lens of nutrition. To illustrate, Hill$_1$ and Hill$_2$ rise as both the richness and evenness of an individual's diet increase, or simply put, when identical quantities of a large number of species are consumed. Whereas richness is advocated in most food-based dietary guidelines worldwide (that is, consume a wide variety of foods *between* and *within* food groups), perfect evenness of food group consumption is in contradiction with all evidence-based dietary recommendations. To illustrate, the EAT–*Lancet* planetary health diet proposes consumption of ~800 kcal per day of whole grains and only ~80 kcal per day of vegetables[38].

However, future research should assess the implications of optimizing the evenness of species intakes on Hill$_1$ and Hill$_2$, in particular. To clarify, ideally, plant species contribute the largest proportion of food consumed (for example, >65% of g per day or kcal per day) (ref. 38), and there is usually a greater diversity of unique plant species in the diet, as compared to animal species[22,23]. Hence, uniform species intakes might in fact proxy food-group-based dietary recommendations. Nonetheless, the effect of weighting factors should be explored for specific food groups (for example, grains vs fruits) or preferably individual edible species to further refine and validate Hill numbers for use in nutritional epidemiology. Notably, using diverging weighting factors (for example, based on an arbitrary selection of either public health or ecological impacts of species) would once again hamper the comparability of food biodiversity indices across studies[21].

We caution against interpreting the absolute values of Hill numbers at face value, due to the absence of references for benchmarking for various dietary assessment methods. At present, the number of effective species should only be used to make relative comparisons between the tails of the scoring distributions (that is, high vs low Hill numbers). In addition, we recognize the substantial variability in chemical components and densities at lower taxonomic levels, such as varieties and cultivars, the impact of production mode, geography (for example, due to soil biodiversity and climate) and seasonality on the nutritional composition of food and the fact that separate edible parts of the same species might belong to distinct functional food groups (for example, pumpkin fruit, seeds and leaves). However, species-level consumption is arguably the most granular taxonomic information that can currently be obtained from self-reported dietary assessments[39].

The current study has several strengths. First, we examined the relationships between the suite of Hill numbers and diet-specific (that is, micronutrient adequacy) and diet-related clinical outcomes (that is, all-cause mortality). Second, the EPIC study's prospective design, large sample size, long and high rates of follow-up and standardized data collection, especially for the usual diet, offers a wide and detailed perspective on food biodiversity in European diets, as reflected by the ~250 unique species reported as consumed. Third, the LMIC study included only dietary assessments that purposely identified food and drinks to at least the species level, as exemplified by the ~230 unique species reported as consumed[23].

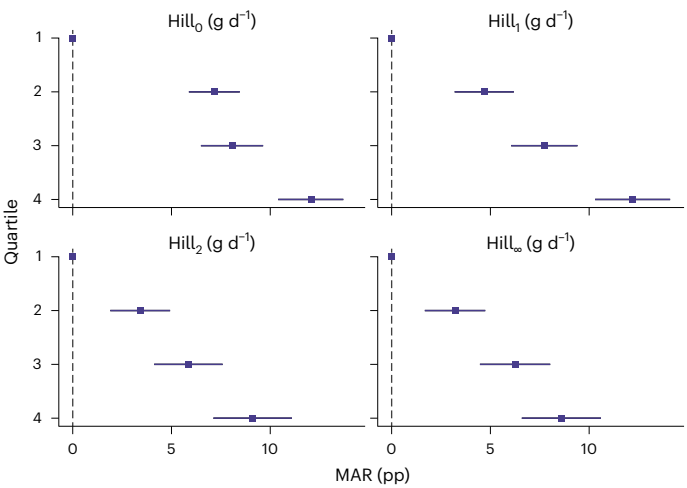

Fig. 5 | Associations between higher quartiles of Hill numbers and micronutrient adequacy in the LMIC study. Linear mixed effects models (random intercept: country; random slope: varying association by country; fixed effect: season and total energy intake) were fitted, assuming an unstructured covariance matrix. Error bars represent the 95% confidence intervals around the point estimate. LMIC, low- and middle-income country (n = 2,186); MAR of calcium, folate, iron, vitamin A, vitamin C and zinc intakes; pp, percentage points. Data availability section provides access details to the source data.

However, our study also has some key limitations that warrant caution. First, the EPIC study enroled potentially more health-conscious, middle-aged volunteers between 1992 and 2000, whereas the LMIC study included only females from rural, subsistence households between 2009 and 2015. Therefore, our findings should not be extrapolated to more contemporary dietary patterns, due to the rapid rise in ultra-processed food consumption[40], other population groups or contexts. Second, both the EPIC and LMIC studies used single-time point assessments of self-reported dietary intake, which are inherently prone to imprecision and inaccuracy and consequently exposure misclassification. The single multiple-pass 24-hour recall method used in the LMIC study allows for estimating population-level intakes but ignores random *within*-person variability and is, as a result, not representative of an individual's usual food (bio)diversity[41]. In contrast, the food frequency questionnaires enumerated in the EPIC study are likely to reflect habitual eating behaviour throughout middle-aged adult life but should only be used to rank individuals as it is unlikely to be an accurate assessment of the absolute number of effective species in diets. The larger absolute $Hill_0$ values in the EPIC study, as compared to the LMIC study, are a methodological artefact of using long-run food frequency questionnaires (that is, different reference periods). It was not possible to transform usual food and drink consumption across one year to average $Hill_0$ over a 24-hour period as, for example, species consumed only twice monthly would falsely be counted as being consumed daily. Conversely, food frequency questionnaires are however an underestimate of habitual $Hill_0$ over the previous year as individuals can only report foods consumed, which were included on the predefined, often constrained, food list. In addition, as details on brand names were not readily available for the EPIC study, the number of species in industrially prepared multi-ingredient items and their contribution to $Hill_0$ is probably underestimated. Third, due to lack of food composition data, our analysis of micronutrient adequacy was constrained to six micronutrients (of public health significance) available across all five LMICs. Furthermore, for a limited number of indigenous, wild, neglected and underutilized species, best-matching food composition values served as proxies[42]. Lastly, our analyses were based on observational studies; therefore, unmeasured, uncontrolled or residual confounding in our regression models cannot be entirely ruled out.

The concept of food biodiversity might help develop localized, cross-sectoral research and actions aimed at achieving global targets, such as those set out in the Sustainable Development Goals, World Health Assembly 2012 and the Kunming–Montreal Global Biodiversity Framework[43]. Furthermore, at national level, the Environment Act 2021 for England included the first legally binding target that total species abundance should be at least 10% higher in 2042 than 2030. Theoretically, higher consumer demand for food biodiversity might contribute to species abundance through a wider range of (commercial) seeds available and accessible to producers and, subsequently, agricultural diversification, which is both an outcome and a potential driver of positive biodiversity impacts[44,45]. Moreover, multiple countries, such as Sweden, Qatar and Uruguay, have included specific dietary principles for biodiversity conservation in their national food-based dietary guidelines[28]. Food biodiversity was shown in this study to be a measurable and valid construct for target setting, which has been a key criticism of previous biodiversity goals (for example, Aichi Targets)[46]. Lastly, Hill numbers calculated from nationally representative dietary intake data might be incorporated into comprehensive tools, such as the Agrobiodiversity Index, to help agrifood systems actors to understand the extent to which their commitments are contributing to the sustainable use and conservation of biodiversity for food and nutrition[1].

In conclusion, $Hill_0$ or H_R has strong construct validity, as exemplified by the positive associations with micronutrient adequacy and negative associations with premature death, has adequate discriminatory power and is comparatively simpler to collect than other Hill numbers, as no intake quantities are required for its calculation. Therefore, as a complement to ongoing assessments of food group consumption (for example, Global Diet Quality Project) and food composition (for example, Periodic Table of Food Initiative) and biodiversity conservation efforts (for example, EU 2030 Biodiversity Strategy), we propose dietary species richness as the most feasible index to monitor and evaluate food biodiversity across contexts and time.

Methods

Our research is reported using the STrengthening the Reporting of OBservational Studies in Epidemiology-Nutritional Epidemiology (STROBE-nut) checklist[47]. Written informed consent was obtained from all study participants. The following analyses were approved by the Ethics Committee of Ghent University Hospital (EC/2019/1015 and NR B670201422403).

Study populations of EPIC and LMIC studies

The EPIC study and the sub-samples of enroled adult females (71%) and males aged 25–70 years at baseline (n = 451,390) from Denmark (n = 55,014), France (n = 67,920), Germany (n = 49,352), Italy (n = 44,547), the Netherlands (n = 36,538), Norway (n = 33,967), Spain (n = 39,990), Sweden (n = 48,690) and the United Kingdom (n = 75,372) included in the present analyses have been extensively described previously[22]. Further details on the EPIC study's data-collection procedures and the identification of foods, drinks and recipes at the species level are provided in Supplementary Note 1.

The five cross-sectional studies, referred to as the LMIC study, and the sub-sample of enroled females aged 18–90 years with dietary intake data (n = 2,149) from the Democratic Republic of Congo (n = 461), Ecuador (n = 251), Kenya (n = 770), Sri Lanka (n = 36) and Viet Nam (n = 631) have been extensively described previously[23]. Further details on the LMIC study's data-collection procedures; the identification of foods, drinks and recipes at the species level and the measure of micronutrient adequacy (that is, MAR of six micronutrients) are provided in Supplementary Note 2.

No specific software was used to collect the food frequency questionnaire data in EPIC or the quantitative 24-hour recall data in the LMIC studies.

Food biodiversity computation using Hill numbers

Hill numbers aim to unify the functional unit, known as the effective number of species, of classical diversity indices (for example, H_R, H_{Sh}, H_{Si} and H_{BP})[26]. The general Hill numbers equation (equation (1)) is as follows:

$$\mathrm{Hill}_a(\mathbf{p}) = \left(\sum_{i=1}^{S} p_i^a\right)^{\frac{1}{1-a}} \quad (1)$$

In equation (1), S represents the total number of distinct edible species in an individual's diet and $\mathbf{p} = (p_1, \ldots, p_S)$ are the proportion p_i of each species i in that individual's diet, or simply put, the weight of a species i divided by the total weight of all food and drinks that contain edible species. The sensitivity parameter a determines the relative weight assigned to rare or common (that is, less or more abundant) species in the diet. In other words, with higher values of a, more weight is placed on species present in large relative quantities (for example, percent of total g per day consumed) in the diet and vice versa. The output of $Hill_a$ is thus the number of effective species, which represents the number of distinct species consumed if the diet were to be perfectly even or when each species is consumed in equal relative quantities[26]. In other words, if an individual's diet has a Hill number equal to x, then this would mean that their diet is as diverse as another person's diet who consumed x unique species in exactly the same quantities (for example, g per day, kcal per day).

We calculated Hill numbers with $a = \{0, 1, 2, \infty\}$, using weight intake (g per day) as the primary quantity variable, to estimate the number of effective species for EPIC and LMIC study participants. In summary, Hill_0 is equivalent to H_R (equation (2)), Hill_1 is the exponential of H_{Sh} (equation (3)), Hill_2 is the reciprocal of H_{Si} (equation (4)) and Hill_∞ is equivalent to H_{BP} (equation (5)) (Fig. 1).

$$\text{Hill}_0(\mathbf{p}) = H_R = S \tag{2}$$

$$\text{Hill}_1(\mathbf{p}) = \exp\left(\sum_{i=1}^{S} p_i \ln(p_i)\right) \tag{3}$$

$$\text{Hill}_2(\mathbf{p}) = \frac{1}{\sum_{i=1}^{S} p_i^2} \tag{4}$$

$$\text{Hill}_\infty(\mathbf{p}) \frac{1}{\max(\mathbf{p})} \tag{5}$$

In equations (2)–(5), $\mathbf{p}$ represents the vector containing all the relative quantities (for example, percent of g per day) of the species in an individual's diet. In a comparative analysis, energy intake (kcal per day), rather than weight intake (g per day), was used as the abundance variable for the Hill numbers computation. Functional and phylogenetic dissimilarity indices were beyond the scope of the manuscript.

Statistical analyses

The distributions of Hill numbers in the EPIC and LMIC studies were visualized using kernel density plots for each individual country (for example, nine countries in the EPIC study) and for each study's pooled sample. Furthermore, the relationships between Hill numbers were assessed using Spearman's rank correlations and visualized using heatscatters (that is, scatterplots with colour-weighted coordinates) for each study's unweighted pooled sample (that is, every individual was considered to be equally representative). If strong to very strong Spearman's rank correlations (that is, $\rho \geq 0.60$) were observed between two Hill numbers, preference would be given to the index with a more straightforward calculation method and lesser data requirements, as this study prioritizes the practical application of indices in nutritional epidemiology, rather than their mathematical behaviour.

Statistical analyses of the EPIC study followed the published plan detailed in the project proposal that was submitted to and approved (March 2019) by the EPIC Steering Committee[22].

Unadjusted absolute mortality rates were calculated as the number of deaths per 10,000 person years in the fifth Q and the first Q of the distribution of each Hill number, respectively. Associations between Hill numbers (that is, Qs of effective species number) and all-cause mortality were characterized using multivariable-adjusted Cox proportional hazard regression models (HR and 95% CI) with age as the primary underlying time variable. The Breslow method was adopted for handling ties. Examination of the Schoenfeld residuals, according to follow-up time (years) for Qs of Hill numbers, confirmed that the proportionality assumptions were satisfied. Participants contributed person time to the model until their date of death, date of emigration/loss to follow-up or end of follow-up, whichever occurred first. *P*-values for assessing linear trends were calculated with the use of the Wald test of a pseudo-continuous score variable, based on the median number of species consumed per year for each Q of a Hill number.

Pooled cohort models were stratified by sex, age at recruitment (1-year intervals) and study centre ('*strata*' option in *proc phreg*, SAS version 9.4 (SAS Institute) and multivariable-adjusted for potential confounding factors using a 5% change-in-estimate criterion for β-coefficients (due to limited knowledge on factors affecting Hill numbers, we tested different socio-demographic, lifestyle and dietary factors potentially associated with either Hill numbers, death rate or both). Subsequently, the plausibility of the causal structure of these variables was cross checked using a directed acyclic graph (Supplementary Fig. 7): smoking status (current, 1 to 15 cigarettes per day; current, 16 to 25 cigarettes per day; current, 26+ cigarettes per day; current, pipe/cigar/occasional; current/former, missing; former, quit 11 to 20 years; former, quit >20 years; former, quit ≤10 years; never; unknown), educational level, as a proxy variable for socio-economic status (longer education (including university degree, technical or professional school); secondary school; primary school completed; not specified), marital status (single, divorced, separated or widowed; married or living together; unknown), physical activity (Cambridge index: active; moderately active; moderately inactive; inactive; missing), alcohol intake at recruitment (g per day), total energy intake (kcal per day), the 18-point relative Mediterranean diet score, as a measure for an overall healthy diet[48], the consumption of red and processed meat (g per day) (ref. 49) and fibre intake (g per day), to reflect carbohydrate quality[50], such as whole grain intake. Notably, our analyses purposely conditioned on the aforementioned dietary risk factors to (conservatively) estimate the independent 'causal effect' of species richness and evenness (that is, Hill numbers). Possible multi-collinearity of (dietary) variables included in our models were assessed by 'collin' and 'vif' options in *proc phreg*, SAS (that is, all condition indices <30 and variance inflation factors <3, respectively). When data on categorical covariates were missing, a 'missing class' was introduced to the model.

To quantify the associations between Hill numbers and the micronutrient adequacy of diets in the LMIC study, a linear mixed effect model was fitted, with random intercepts for countries, random slopes for the heterogeneous associations by country and the season of data collection (lean, plenty) and total energy intake (kcal per day) as fixed effect variables, assuming an unstructured covariance matrix. The normality of residuals were assessed by visual inspection of the normal probability and quantile-normal plots. If the normality assumption had been violated, Hill numbers were to be transformed or quantile regression models were to be fitted. These analyses were not replicated in the EPIC study, as long-run food frequency questionnaires, by design, do not provide accurate estimates of absolute intakes; hence, estimated micronutrient intakes cannot be validly compared to their requirements.

The relationships between Qs and quartiles of Hill numbers and HR for all-cause mortality in the EPIC study and MAR in the LMIC study, respectively, were visualized using interval plots for each study's pooled sample.

All statistical tests were two sided and $P < 0.05$ was considered statistically significant. Our exploratory analyses did not correct for multiple comparisons to avoid inflating type II error (that is, false negative findings). SAS software version 9.4 (SAS Institute) was used for the statistical analyses and R software version 4.2.3 (RStudio) was used to calculate Hill numbers (that is, hillR package) and for data visualization.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

EPIC data and biospecimens are available to investigators in the context of research projects that are consistent with the legal and ethical standard practices of IARC/WHO and the EPIC Centers. The use of a random sample of anonymized data from the EPIC study can be requested by contacting epic@iarc.fr. Information on the EPIC data-access policy and on how to submit an application for gaining access to EPIC data and/or biospecimens is available at http://epic.iarc.fr/access/index.php. The cross-sectional studies from the Democratic Republic of Congo, Ecuador, Kenya and Viet Nam were approved by an ethics committee, while in Sri Lanka, the protocol was exempted from clearance. The

anonymized individual-level data and their protocols are available at https://dataverse.harvard.edu/dataverse/DietarySpeciesRichness. Source data are provided with this paper.

Code availability

Analytical code is available via Zenodo at https://doi.org/10.5281/zenodo.14730448 (ref. 51).

References

1. Jones, S. K. et al. Agrobiodiversity Index scores show agrobiodiversity is underutilized in national food systems. *Nat. Food* **2**, 712–723 (2021).
2. Benton, T., Bieg, C., Harwatt, H., Pudassaini, R. & Wellesley, L. *Food System Impacts on Biodiversity Loss: Three Levers for Food System Transformation in Support of Nature* (Chatham House, The Royal Institute of International Affairs, 2021).
3. Khoury, C. K. et al. Increasing homogeneity in global food supplies and the implications for food security. *Proc. Natl Acad. Sci. USA* **111**, 4001–4006 (2014).
4. Verger, E. O. et al. Dietary diversity indicators and their associations with dietary adequacy and health outcomes: a systematic scoping review. *Adv. Nutr.* **12**, 1659–1672 (2021).
5. Stevens, G. A. et al. Micronutrient deficiencies among preschool-aged children and women of reproductive age worldwide: a pooled analysis of individual-level data from population-representative surveys. *Lancet Glob. Health* **10**, e1590–e1599 (2022).
6. Estrada-Carmona, N., Sánchez, A. C., Remans, R. & Jones, S. K. Complex agricultural landscapes host more biodiversity than simple ones: a global meta-analysis. *Proc. Natl Acad. Sci. USA* **119**, e2203385119 (2022).
7. Herrero, M. et al. Farming and the geography of nutrient production for human use: a transdisciplinary analysis. *Lancet Planet. Health* **1**, e33–e42 (2017).
8. Clark, M. A. et al. Global food system emissions could preclude achieving the 1.5° and 2 °C climate change targets. *Science* **370**, 705–708 (2020).
9. Remans, R., Wood, S. A., Saha, N., Anderman, T. L. & DeFries, R. S. Measuring nutritional diversity of national food supplies. *Glob. Food Sec.* **3**, 174–182 (2014).
10. Thompson, H. J. et al. Dietary botanical diversity affects the reduction of oxidative biomarkers in women due to high vegetable and fruit intake. *J. Nutr.* **136**, 2207–2212 (2006).
11. DeClerck, F., Fanzo, J., Palm, C. & Remans, R. Ecological approaches to human nutrition. *Food Nutr. Bull.* **32**, S41–50 (2011).
12. Springmann, M., Godfray, H. C. J., Rayner, M. & Scarborough, P. Analysis and valuation of the health and climate change cobenefits of dietary change. *Proc. Natl Acad. Sci.* **113**, 4146–4151 (2016).
13. Jacobs, D. R. & Tapsell, L. C. Food, not nutrients, is the fundamental unit in nutrition. *Nutr. Rev.* **65**, 439–450 (2007).
14. Herforth, A. W., Wiesmann, D., Martínez-Steele, E., Andrade, G. & Monteiro, C. A. Introducing a suite of low-burden diet quality indicators that reflect healthy diet patterns at population level. *Curr. Dev. Nutr.* **4**, nzaa168 (2020).
15. Barabási, A. L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nat. Food* **1**, 33–37 (2020).
16. Eisenhauer, N. et al. Plant diversity effects on soil food webs are stronger than those of elevated CO_2 and N deposition in a long-term grassland experiment. *Proc. Natl Acad. Sci. USA* **110**, 6889–6894 (2013).
17. de Oliveira Otto, M. C. et al. Dietary diversity: implications for obesity prevention in adult populations: a science advisory from the American Heart Association. *Circulation* **138**, e160–e168 (2018).
18. Medeiros de, M. F. A. et al. Assessment of biodiversity in food consumption studies: a systematic review. *Front. Nutr.* **9**, 832288 (2022).
19. Berti, P. R. Relationship between production diversity and dietary diversity depends on how number of foods is counted. *Proc. Natl Acad. Sci.* **112**, e5656 (2015).
20. Daly, A. J., Baetens, J. M. & De Baets, B. Ecological diversity: measuring the nmeasurable. *Mathematics* **6**, 119 (2018).
21. Hanley-Cook, G. T. et al. Food biodiversity: quantifying the unquantifiable in human diets. *Crit. Rev. Food Sci. Nutr*. https://doi.org/10.1080/10408398.2022.2051163 (2022).
22. Hanley-Cook, G. T. et al. Food biodiversity and total and cause-specific mortality in 9 European countries: an analysis of a prospective cohort study. *PLoS Med.* **18**, e1003834 (2021).
23. Lachat, C. et al. Dietary species richness as a measure of food biodiversity and nutritional quality of diets. *Proc. Natl Acad. Sci.* **115**, 127–132 (2018).
24. Nelson, G. et al. Income growth and climate change effects on global nutrition security to mid-century. *Nat. Sustain.* **1**, 773–781 (2018).
25. Salomé, M. et al. Contrary to ultra-processed foods, the consumption of unprocessed or minimally processed foods is associated with favorable patterns of protein intake, diet quality and lower cardiometabolic risk in French adults (INCA3). *Eur. J. Nutr.* **60**, 4055–4067 (2021).
26. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
27. Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).
28. James-Martin, G. et al. Environmental sustainability in national food-based dietary guidelines: a global review. *Lancet Planet. Health* **6**, e977–e986 (2022).
29. Roswell, M., Dushoff, J. & Winfree, R. A conceptual guide to measuring species diversity. *Oikos* **130**, 321–338 (2021).
30. Fang, C., Fernie, A. R. & Luo, J. Exploring the diversity of plant metabolism. *Trends Plant Sci.* **24**, 83–98 (2019).
31. Ahmed, S. et al. Foodomics: a data-driven approach to revolutionize nutrition and sustainable diets. *Front. Nutr.* **9**, 874312 (2022).
32. Bernhardt, J. R. & O'Connor, M. I. Aquatic biodiversity enhances multiple nutritional benefits to humans. *Proc. Natl Acad. Sci.* **118**, e1917487118 (2021).
33. Heilpern, S. A. et al. Declining diversity of wild-caught species puts dietary nutrient supplies at risk. *Sci. Adv.* **7**, eabf9967 (2021).
34. Heiman, M. L. & Greenway, F. L. A healthy gastrointestinal microbiome is dependent on dietary diversity. *Mol. Metab.* **5**, 317–320 (2016).
35. Valdes, A. M., Walter, J., Segal, E. & Spector, T. D. Role of the gut microbiota in nutrition and health. *BMJ* **361**, k2179 (2018).
36. Xiao, C. et al. Associations of dietary diversity with the gut microbiome, fecal metabolites, and host metabolism: results from 2 prospective Chinese cohorts. *Am. J. Clin. Nutr.* **116**, 1049–1058 (2022).
37. Fardet, A. & Rock, E. Chronic diseases are first associated with the degradation and artificialization of food matrices rather than with food composition: calorie quality matters more than calorie quantity. *Eur. J. Nutr.* **61**, 2239–2253 (2022).
38. Willett, W. C. et al. Food in the Anthropocene: the EAT–*Lancet* Commission on healthy diets from sustainable food systems. *Lancet* **6736**, 3–49 (2019).
39. Food and Agriculture Organization of the United Nations & Bioversity International *Guidelines on Assessing Biodiverse Foods in Dietary Intake Surveys* (FAO, 2017).
40. Leite, F. H. M. et al. Ultra-processed foods should be central to global food systems dialogue and action on biodiversity. *BMJ Glob. Health* **7**, e008269 (2022).

41. Hanley-Cook, G. T. et al. Seasonality and day-to-day variability of dietary diversity: longitudinal study of pregnant women enrolled in a randomized controlled efficacy trial in rural Burkina Faso. *J. Nutr.* **152**, 2145–2154 (2022).
42. Charrondière, U. R. et al. FAO/INFOODS food composition database for biodiversity. *Food Chem.* **140**, 408–412 (2013).
43. World Health Organization *Guidance on Mainstreaming Biodiversity for Nutrition and Health* (WHO, 2020).
44. Hanley-Cook, G. T., Kennedy, G. & Lachat, C. *Reducing Risk of Poor Diet Quality through Food Biodiversity. Agrobiodiversity Index Report* (Bioversity International, 2019); https://doi.org/10.1016/b978-0-323-01199-0.50226-7
45. Jones, S. K. et al. Achieving win-win outcomes for biodiversity and yield through diversified farming. *Basic Appl. Ecol.* **67**, 14–31 (2023).
46. Green, E. J. et al. Relating characteristics of global biodiversity targets to reported progress. *Conserv. Biol.* **33**, 1360–1369 (2019).
47. Lachat, C. et al. Strengthening the reporting of observational studies in epidemiology—nutritional epidemiology (STROBE-nut): an extension of the STROBE statement. *PLoS Med.* **13**, e1002036 (2016).
48. Buckland, G. et al. Adherence to the mediterranean diet and risk of coronary heart disease in the Spanish EPIC cohort study. *Am. J. Epidemiol.* **170**, 1518–1529 (2009).
49. Zheng, Y. et al. Association of changes in red meat consumption with total and cause specific mortality among US women and men: two prospective cohort studies. *BMJ* **365**, l2110 (2019).
50. Reynolds, A. et al. Carbohydrate quality and human health: a series of systematic reviews and meta analyses. *Lancet* **393**, 434–445 (2019).
51. Deygers, J., Berden, J. & Hanley-Cook, G. T. Syntax: unifying food biodiversity indices through effective species numbers: analysis of cohorts in Europe and surveys from Africa, Asia, and South America. *Zenodo* https://doi.org/10.5281/zenodo.14730448 (2025).

Acknowledgements

We are thankful for the EPIC-Ragusa data provided by the Hyblean Association for Epidemiology Research (AIRE-ONLUS). The coordination of EPIC is financially supported by the International Agency for Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are supported by the Danish Cancer Society (Denmark); Ligue Nationale Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale (MGEN), Institut National de la Santé et de la Recherche Médicale (INSERM), French National Research Agency (ANR, reference ANR-10-COHO-0006), French Ministry for Higher Education (subsidies 2102918823, 2103236497 and 2103586016) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE) and Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro (AIRC), Compagnia di SanPaolo, and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF) and Statistics Netherlands (The Netherlands); Health Research Fund (FIS)–Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra and the Catalan Institute of Oncology (ICO) (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (C864/A14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford) and Medical Research Council (MR/N003284/1, MC-UU_12015/1 and MC_UU_00006/1 to EPIC-Norfolk; MR/M012190/1 to EPIC-Oxford) (United Kingdom). The cross-sectional studies in Africa, Asia and South America were supported by the Flemish Interuniversity Council (Ecuador); Flemish Interuniversity Council, Leopold III fund for Nature Exploration and Conservation and Stichting Roeping (Democratic Republic of Congo); Global Environment Facility, United Nations Environmental Programme, Food and Agriculture Organization of the United Nations and Bioversity International (Sri Lanka); and Humidtropics and A4NH (Kenya and Viet Nam). Funding for grant number IIG_FULL_2020_034 was obtained from Wereld Kanker Onderzoek Fonds (WKOF) as part of the World Cancer Research Fund International grant programme (principal investigator: I.H.; co-investigators: G.T.H.-C., P.V., K.A.M., M.D.-T., E.K.-G., M.T. and C.L.). D.B.I. was supported by a grant from the Independent Research Fund Denmark (1057-00016B) and the Danish Diabetes Association. P.V. was supported by the NIHR Global Health Research Centres on non-communicable diseases and Environmental Change (NIHR203247). Researchers were independent from the funders. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Where authors are identified as personnel of IARC/World Health Organization (WHO), the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy or views of IARC/WHO.

Author contributions

G.T.H.-C., A.J.D. and C.L. designed the study; G.T.H.-C., J.D., J. Berden and I.H. conducted the research; J.D., J. Berden and G.T.H.-C. analysed data and performed statistical analysis; G.T.H.-C. provided support for the statistical analysis; G.T.H.-C. and J.D. developed the first draft and revised the manuscript: A.J.D., J. Berden, R.R., C.T., D.B.I., J. Baudry, P.V.D., E.K.-G., I.H. and C.L. critically reviewed the manuscript; and P.V., M.B.S., K.T.H., M.D.-T., A.H., C.C.D., Y.v.d.S., G.S., M.G., L.M., D.P., J.E.R., F.O.O., D.H., D.R., K.A.M. and M.T. read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43016-025-01147-6.

Correspondence and requests for materials should be addressed to Giles T. Hanley-Cook or Carl Lachat.

Peer review information *Nature Food* thanks Sebastian Heilpern, Fernanda Helena Marrocos Leite, Michelle Cristine Medeiros Jacob and Jörg Müller for their contribution to the peer review of this work.

[1]Department of Food Technology, Safety and Health, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium. [2]Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium. [3]Nutrition and Metabolism Branch, International Agency for Research on Cancer, Lyon, France. [4]The Alliance of Bioversity International and the International Center for Tropical Agriculture, Montpellier, France. [5]Alliance of Bioversity International and the International Center for Tropical Agriculture, Nairobi, Kenya. [6]Department of Public Health, Aarhus University, Aarhus, Denmark. [7]Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark. [8]Department of Nutrition, Sports and Exercise, University of Copenhagen, Frederiksberg, Denmark. [9]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK. [10]Sorbonne Paris Nord University, Inserm, INRAE, Cnam, Nutritional Epidemiology Research Team (EREN), Epidemiology and Statistics Research Center-Paris Cité University (CRESS), Bobigny, France. [11]Faculty of Tropical AgriSciences, Czech University of Life Sciences, Prague, Czech Republic. [12]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [13]Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany. [14]Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany. [15]Alliance of Bioversity International and the International Center for Tropical Agriculture, Hanoi, Viet Nam. [16]Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, UK. [17]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands. [18]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway. [19]Instituto de Salud Pública y Laboral de Navarra, Pamplona, Spain. [20]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [21]Navarra Institute for Health Research (IdiSNA), Pamplona, Spain. [22]Centre for Biostatistics, Epidemiology, and Public Health, Department of Clinical and Biological Sciences, University of Turin, Turin, Italy. [23]Faculty of Medicine, Universidad de Especialidades Espíritu Santo, Samborondon, Ecuador. [24]Department of Food Science, Nutrition and Technology, Faculty of Agriculture, University of Nairobi, Nairobi, Kenya. [25]Alliance of Bioversity International and the International Center for Tropical Agriculture, Rome, Italy. [26]Department of Agricultural Biology, Faculty of Agriculture, University of Ruhuna, Matara, Sri Lanka. [27]Medical Research Council Unit The Gambia, London School of Hygiene and Tropical Medicine, London, UK. [28]These authors contributed equally: Giles T. Hanley-Cook, Jill Deygers. ✉e-mail: Giles.HanleyCook@UGent.be; Carl.Lachat@UGent.be

nature portfolio

Corresponding author(s): Giles T. Hanley-Cook
Carl K. Lachat

Last updated by author(s): Jan 23, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specific software was used to collect the food frequency questionnaire data in EPIC or the quantitative 24-hour recall data in the LMIC studies.
Data analysis	SAS software, version 9.4 (SAS Institute) was used for the statistical analyses and R software, version 4.2.3 (RStudio) was used to calculate Hill numbers (i.e., hillR package) and for data visualization. Analytical code is available at https://doi.org/10.5281/zenodo.14730448

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

EPIC data and biospecimens are available to investigators in the context of research projects that are consistent with the legal and ethical standard practices of IARC/WHO and the EPIC Centers. The use of a random sample of anonymized data from the EPIC study can be requested by contacting epic@iarc.fr. For information

on EPIC data access policy and on how to submit an application for gaining access to EPIC data and/or biospecimens, please follow the instructions at http://epic.iarc.fr/access/index.php.
The cross-sectional studies from the Democratic Republic of Congo, Ecuador, Kenya, and Viet Nam were approved by an ethics committee, while in Sri Lanka the protocol was exempted from clearance. The anonymized individual-level data and their protocols are available at https://dataverse.harvard.edu/dataverse/DietarySpeciesRichness.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	In both the EPIC and LMIC studies, sex (female, male) was self-reported. In the EPIC study, multi-adjusted models were stratified for centre, age at recruitment (1-year intervals, timescale), and sex and adjusted for baseline alcohol intake (g/day), physical activity (Cambridge index: active; moderately active; moderately inactive; inactive; missing), marital status (single, divorced, separated, or widowed; married or living together; unknown), smoking status and intensity of smoking (current, 1 to 15 cigarettes/day; current, 16 to 25 cigarettes/day; current, 26+ cigarettes/day; current, pipe/cigar/occasional; current/former, missing; former, quit 11 to 20 years; former, quit >20 years; former, quit ≤10 years; never; unknown), educational level [longer education (including university degree, technical or professional school); secondary school; primary school completed; not specified], baseline energy intake (kcal/day), baseline fiber intake (g/day), baseline red and processed meat consumption (g/day), and an 18-point Mediterranean diet score. In the LMIC study, for each female, we calculated the mean adequacy ratio as the arithmetic mean of the quantity of a nutrient consumed on a daily bases per its age-, sex-, and physiological status-specific Estimated Average Requirement (i.e., the cut-point covering the requirements of 50% of the population).
Population characteristics	In the EPIC study, an extensive and standardized phenotypic characterization was performed for each participant upon enrolment. Questionnaires were used to collect sociodemographic information, educational level (standardized for the whole cohort), personal and familial history of diseases, lifestyle (e.g., smoking, alcohol consumption, and physical activity), and menstrual and reproductive history of females. Anthropometric measurements (e.g., height, weight, waist, and hip circumferences) were performed in all centres (except France, Oxford, and Norway: self-reported data). Height and weight were complemented with available self-reported values or imputed with centre-, age-, and gender-specific average values when missing. Body mass index (BMI) was calculated as weight divided by height squared (kg/m^2). Usual dietary intake was assessed for each individual at recruitment using country- or centre-specific validated dietary questionnaires (DQs) developed to capture the geographical specificity of an individual's diet over the preceding year. The type of DQ used differed according to study centres and included self- or interviewer-administered semi-quantitative food frequency questionnaires (FFQs) with an estimation of individual average portions or with the same standard portion assigned to all participants, or diet history questionnaires combining an FFQ and 7-day dietary records. In most centres, DQs were self-administered, with the exception of Ragusa, Naples, and Spain, where face-to-face interviews were conducted. Extensive quantitative DQs were used in northern Italy, the Netherlands, and Germany, which were structured by meals in Spain, France, and Ragusa. Semi-quantitative FFQs were used in Denmark, Norway, Naples, Umeå, and the UK, while an FFQ was combined with a 7-day record on hot meals in Malmö. Data on vital status and cause and date of death were obtained using record linkages with population-based cancer registries, boards of health, health insurance registries, pathology registries and mortality registries (Denmark, Italy, the Netherlands, Norway, Spain, Sweden, and the UK), or through active follow-up and next of kin (France and Germany). Germany identified deceased individuals from undelivered follow-up mailings and subsequent enquiries to municipality registries, regional health departments, physicians, or hospitals. In France, information on deceased participants was obtained using the database of health insurance for school employees and national death index. The end of follow-up/closure dates for death ascertainment varied between 2009 and 2014 depending on the countries. In the LMIC study, questionnaires were used to collect information on sex, age, and physiological status. Dietary intake data of females were collected using a quantitative and comparable single multiple-pass 24-h dietary recall method. Foods and drinks, including mixed dishes, were identified at the species-level following best-practice guidelines using local taxonomical references and technical expertise.
Recruitment	The EPIC study (http://epic.iarc.fr/) is an ongoing multicentre, prospective cohort study investigating metabolic, dietary, lifestyle, and environmental factors in relation to cancer and other chronic diseases. Between 1992 and 2000, more than 500,000 volunteers (aged 25 to 70 years) were recruited from 10 European countries (23 administrative centres): Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom. Most of the participants were recruited from the general population residing in a given geographic area, town, or province. Exceptions were the cohorts of France (female members of a health insurance scheme for school employees), Utrecht (breast cancer screening attendees), Ragusa (blood donors and their spouses), and Oxford (mainly vegetarian and healthy eaters). All participants gave written informed consent and completed questionnaires on their diet, lifestyle, and medical history. The LMIC study data were collected among females aged 12-90 years (n=2,186) during the wet (lean) season from rural areas in the Democratic Republic of Congo (n=461),11 Ecuador (n=257),12 Kenya (n=394),13 Sri Lanka (n=36) and Viet Nam (n=268). Dietary intake data from the dry (plenty) season were also available from Kenya (n=396)13 and Viet Nam (n=374). All studies were conducted between July 2009 and April 2015, and samples were representative of the village-level populations.
Ethics oversight	The analyses for this study were approved by the Ethics Committee of Ghent University Hospital (EC/2019/1015 and NR B670201422403).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Data are (semi-)quantitative and were obtained from (i) prospective cohort studies in nine European countries, including 451,390 adult females and males and (ii) cross-sectional studies in five countries in the Global South, including 2,149 females.
Research sample	The present study included a sub-sample of adults enrolled in the EPIC study (n=451,390) from Denmark (n=55,014), France (n=67,920), Germany (n=49,352), Italy (n=44,547), the Netherlands (n=36,538), Norway (n=33,967), Spain (n=39,990), Sweden (n=48,690), and the United Kingdom (n=75,372). The mean (SD) age of participants at enrolment was 51 years (10), 71% were females, 60% were married or living together, 16% were single, divorced, separated or widowed, 28% received no education or completed primary school only, 24% received longer education (including university degrees), 23% attended a technical/professional school only, 21% attended secondary school only, 49% never smoked, 27% were former smokers, 22% were current smokers, according to the Cambridge index, 33% were moderately inactive, 27% were moderately active, 20% were inactive, and 19% were active, mean BMI was 25.3 kg/m² (4.2), mean height was 166 cm (9), mean weight was 70.0 kg (13.6), 9% had a family history of breast cancer, and 9% had a family history of colorectal cancer. The nine prospective cohorts are not nationally representative. Furthermore, the present study also included a sub-sample of females from five cross-sectional studies (n=2,149) in the Democratic Republic of Congo (DRC) (n=461), Ecuador (n=251), Kenya (n=770), Sri Lanka (n=36), and Viet Nam (n=631). The median (P25, P75) age of females was 35 years (27, 45) in the DRC, 38 years (29, 47) in Ecuador, 48 years (35, 58) in Sri Lanka, 23 years (21, 26) years in both the wet and dry seasons in Viet Nam, 27 years (23, 35) years in the wet season in Kenya, and 27 (23, 36) years in the dry season in Kenya. Samples were representative of the village-level populations, but not representative of all females. The EPIC and LMIC study data were purposively selected. In the EPIC study, foods and drinks were previously identified at the species-level using the European Food Safety Authority's FoodEx2 food classification and description system in combination with the detailed EPIC food classification system (NCLASS). For all countries, composite dishes were decomposed into their ingredients (species) using standard recipes. In the LMIC study, foods and drinks, including mixed dishes, were identified at the species-level following best-practice guidelines using local taxonomical references and technical expertise. Studies with an ethnobotanical focus, had thorough documentation of species verified through a local herbarium, botanical museum, botanist or biologist. Only data from Viet Nam and Sri Lanka identified variety- or cultivar-level where possible. When it was not possible to identify the specific species, the identified genus was followed by sp. Given specific disambiguation challenges, all bananas were recorded as Musa sp. Disambiguation of scientific names was performed using published databases (The Plant List: www.theplantlist.org/; Species 2000 & ITIS Catalogue of Life: www.catalogueoflife.org/col). The LMICS study considered agricultural and wild sources of food and drinks, without assessment of dietary supplement intake.
Sampling strategy	The EPIC study (http://epic.iarc.fr/) is an ongoing multicentre, prospective cohort study investigating metabolic, dietary, lifestyle, and environmental factors in relation to cancer and other chronic diseases. Between 1992 and 2000, more than 500,000 volunteers (aged 25 to 70 years) were recruited from 10 European countries (23 administrative centres): Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom. Most of the participants were recruited from the general population residing in a given geographic area, town, or province. Exceptions were the cohorts of France (female members of a health insurance scheme for school employees), Utrecht (breast cancer screening attendees), Ragusa (blood donors and their spouses), and Oxford (mainly vegetarian and healthy eaters). Data from the the LMIC study was either from a convenient sample of females or a random sample drawn from a village census.
Data collection	In the EPIC study, an extensive and standardized phenotypic characterization was performed for each participant upon enrolment. Questionnaires were used to collect sociodemographic information, educational level (standardized for the whole cohort), personal and familial history of diseases, lifestyle (e.g., smoking, alcohol consumption, and physical activity), and menstrual and reproductive history of females. Anthropometric measurements (e.g., height, weight, waist, and hip circumferences) were performed in all centres (except France, Oxford, and Norway: self-reported data). Height and weight were complemented with available self-reported values or imputed with centre-, age-, and gender-specific average values when missing. Body mass index (BMI) was calculated as weight divided by height squared (kg/m²). Usual dietary intake was assessed for each individual at recruitment using country- or centre-specific validated dietary questionnaires (DQs) developed to capture the geographical specificity of an individual's diet over the preceding year. The type of DQ used differed according to study centres and included self- or interviewer-administered semi-quantitative food frequency questionnaires (FFQs) with an estimation of individual average portions or with the same standard portion assigned to all participants, or diet history questionnaires combining an FFQ and 7-day dietary records. In most centres, DQs were self-administered, with the exception of Ragusa, Naples, and Spain, where face-to-face interviews were conducted. Extensive quantitative DQs were used in northern Italy, the Netherlands, and Germany, which were structured by meals in Spain, France, and Ragusa. Semi-quantitative FFQs were used in Denmark, Norway, Naples, Umeå, and the UK, while an FFQ was combined with a 7-day record on hot meals in Malmö. Data on vital status and cause and date of death were obtained using record linkages with population-based cancer registries, boards of health, health insurance registries, pathology registries and mortality registries (Denmark, Italy, the Netherlands, Norway, Spain, Sweden, and the UK), or through active follow-up and next of kin (France and Germany). Germany identified deceased individuals from undelivered follow-up mailings and subsequent enquiries to municipality registries, regional health departments, physicians, or hospitals. In France, information on deceased participants was obtained using the database of health insurance for school employees and national death index. The end of follow-up/closure dates for death ascertainment varied

between 2009 and 2014 depending on the countries. There was no blinding of participants or researchers.
In the LMIC study, questionnaires were used to collect information on sex, age, and physiological status. Dietary intake data of females were collected using a quantitative and comparable interviewer-administered single multiple-pass 24-h dietary recall method. Foods and drinks, including mixed dishes, were identified at the species-level following best-practice guidelines using local taxonomical references and technical expertise. There was no blinding of participants or researchers.
The present analysis is a secondary analysis of data, hence the research hypothesis was unknown by the participants and researchers at enrolment.

Timing

The EPIC study (http://epic.iarc.fr/) is an ongoing multicentre, prospective cohort study investigating metabolic, dietary, lifestyle, and environmental factors in relation to cancer and other chronic diseases. Between 1992 and 2000, more than 500,000 volunteers (aged 25 to 70 years) were recruited from 10 European countries (23 administrative centres): Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom. Questionnaires, anthropometric measurements, and usual dietary intake was assessed for each individual at recruitment.The end of follow-up/closure dates for death ascertainment varied between 2009 and 2014 depending on the countries.
The LMIC study data were collected among females aged 12-90 years (n=2,186) during the wet (lean) season from rural areas in the Democratic Republic of Congo (n=461), Ecuador (n=257), Kenya (n=394), Sri Lanka (n=36) and Viet Nam (n=268). Dietary intake data from the dry (plenty) season were also available from Kenya (n=396) and Viet Nam (n=374). All studies were conducted between July 2009 and April 2015, and samples were representative of the village-level populations.

Data exclusions

In the EPIC study, we excluded participants with missing dietary information (n=6,342), those with an extreme ratio of energy intake to energy requirement (top and bottom 1%, as these values were considered physiologically implausible) (n=9,607), volunteers with null follow-up (n=1,864), those with prevalent disease at baseline (history of cancer, cardiovascular diseases, and diabetes) (n=23,560), and all participants from the EPIC-Greece cohort (n=28,561), due to administrative constraints.
In the LMIC study, participants with missing dietary intake data (n=37) were excluded. In the event food composition data were missing, best-matching values were obtained from similar settings, countries, or food and drinks.

Non-participation

Following regulations for medical research on humans, data were only collected from participants that provided written informed consent in the EPIC and LMIC studies. No data were collected from participants that refused to participate. No information is available on the non-participation rate.

Randomization

The EPIC and LMIC studies were non-experimental, observational studies.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging