NaturalResourcesfor Human Health

eISSN: 2583-1194

Original Research

Received 16 April 2026
Revised 20 May 2026
Accepted 22 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 1068–1085

KEYWORDS:
Vitis vinifera; polyphenols; dietary exposure assessment; data integration; cloud computing; prospective cohort.

Cloud-Based Integration of Vitis Vinifera Phenolic and Cohort Data Questions Beverage-Type Polyphenol Proxies

Deepak Yadav[1]*, Savita Kumari Sheoran[1]

[1]Department of Computer Science and Engineering, Indira Gandhi University, Meerpur, Rewari 122502, Haryana, India.
ORCID: Deepak Yadav, 0009-0005-1708-9042; Savita Kumari Sheoran, 0000-0002-0359-4700.
Savita.sheoran@igu.ac.in
*Corresponding author: Deepak Yadav, Department of Computer Science and Engineering, Indira Gandhi University, Meerpur, Rewari 122502, Haryana, India.
E-mail: Deepak.cse.rs@igu.ac.in

ABSTRACT:

Background. Research on plant-derived bioactives routinely substitutes the food or beverage that carries a compound for the compound itself, and the heterogeneous data needed to test that substitution are rarely brought into a single queryable structure.

Methods. Two openly available sources were integrated in a containerised, cloud-executed pipeline: a phenolic analysis of 178 wine samples from three Vitis vinifera cultivars, and the NHANES I Epidemiologic Follow-up Study cohort of 1,629 adults. A six-table star schema was loaded into a row store and a columnar store, and query latency, data-volume scaling and parallel execution were benchmarked. Composition was compared by analysis of variance and principal component analysis; 10-year all-cause mortality was modelled by sequentially adjusted logistic regression, four classifiers, feature ablation and E-value analysis.

Results. The columnar store was 34 times faster than the row store at 1.63 million fact rows, and the single-threaded analytical workload reached 79% parallel efficiency on two cores. Cultivar explained 24–73% of variance in phenolic constituents, with a 3.8-fold difference in flavanoid content. Wine drinkers had lower crude mortality than beer drinkers (9.8% versus 18.3%; odds ratio 0.43, 95% CI 0.24–0.77), but adjustment removed 58% of the association (0.70, 0.35–1.39; E-value 1.68) and removing beverage type changed discrimination by under 0.001 AUC.

Conclusion. Beverage type is a poor surrogate for polyphenol dose. Natural-product epidemiology should quantify the bioactive, and the integration infrastructure to do so is inexpensive.

1. INTRODUCTION

Plant polyphenols are among the most intensively studied natural products in human nutrition. Flavanoids, proanthocyanidins, phenolic acids and stilbenes are proposed to act through antioxidant, anti-inflammatory, endothelial and microbiota-mediated pathways, and prospective cohorts consistently report inverse associations between habitual flavonoid intake and all-cause mortality (Bondonno et al., 2019; Grosso et al., 2017; Parmenter et al., 2025). Mechanistic and

pharmacokinetic work has clarified which structures reach the circulation, in what form and at what concentration (Del Rio et al., 2013; Manach et al., 2005).

A methodological gap separates these two literatures. Chemical studies quantify a defined compound in a defined matrix; epidemiological studies almost never measure the compound at all. Instead they record how much of a carrier food or beverage a participant consumes and treat that as a proxy for exposure to the bioactive it contains. The proxy is only as good as two assumptions: that the carrier has a reasonably stable content of the compound, and that consumption of the carrier is not itself a marker for something else that affects the outcome.

Wine is the archetype. Following the report that wine consumption might explain low coronary mortality in France despite a high saturated-fat intake (Renaud and de Lorgeril, 1992), grape polyphenols became the presumed active principle, and 'wine drinking' became the exposure of record in dozens of cohort analyses. Yet wine is a chemically heterogeneous product. Phenolic content depends on cultivar, ripeness, viticultural region, vintage and vinification technique (El Rayess et al., 2024; Waterhouse, 2002), and compositional databases record wide ranges rather than single values for nominally identical products (Neveu et al., 2010). At the same time, the epidemiological literature on alcohol has been substantially revised: genetic and meta-analytic evidence now attributes much of the apparent benefit of moderate drinking to residual confounding and to misclassification of former drinkers as abstainers (Millwood et al., 2019; Zhao et al., 2023).

Testing the two assumptions together is as much a data-management problem as a scientific one. Compositional chemistry and population epidemiology are held in different formats, at different grains and under different governance regimes, and joining them has historically meant ad hoc scripts whose provenance is difficult to audit. Modern practice in cloud data management offers a direct remedy: a normalised analytical schema, a storage engine matched to the query pattern, containerised execution with an explicit resource allocation, and machine-readable provenance (Abadi et al., 2013; Raasveldt and Mühleisen, 2019). These techniques are standard in data-intensive computing but are not yet routine in natural-product research, where the analytical unit is often a single spreadsheet.

This study evaluates both assumptions on one such platform. We first quantify how much the phenolic composition of wine varies with cultivar alone, holding growing region constant, using a classical analytical data set of 178 samples from three *Vitis vinifera* cultivars. We then estimate the association between preferred alcoholic beverage and 10-year all-cause mortality in a national cohort, tracking how that association behaves as confounders are added and testing whether beverage type carries any independent predictive information. The contributions of this work are: (i) an integrated star schema and containerised pipeline for joining phytochemical composition with cohort outcomes, benchmarked across two storage engines, four data volumes and three degrees of parallelism; (ii) a quantification of between-cultivar variance in seven phenolic and colour constituents; (iii) a sequentially adjusted analysis of beverage preference and mortality in the NHANES I Epidemiologic Follow-up Study, with feature ablation and quantitative bias analysis; and (iv) practical guidance on exposure measurement and on provisioning the computation that supports it.

2. MATERIALS AND METHODS

2.1 Study design

This is a secondary analysis of two independent, openly available data sets, organised as two parallel strands that are interpreted jointly and that execute on a common data-management platform (Figure 1). Strand A characterises the phenolic composition of the exposure vehicle. Strand B estimates the health association attributed to that vehicle. No new human or animal data were collected.

NaturalResourcesfor
Human Health
eISSN: 2583-1194

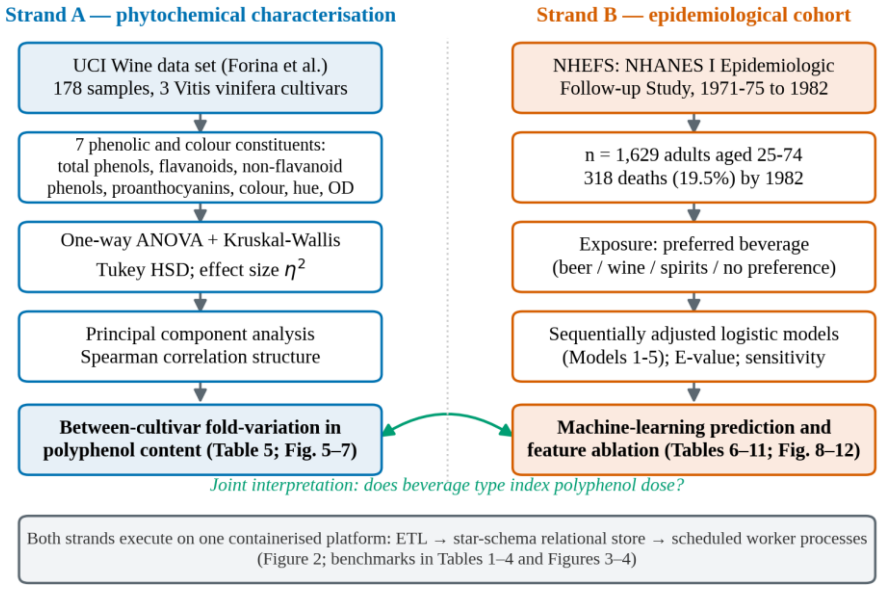

Figure 1. Study design and analytical flow. Strand A characterises the phenolic composition of the exposure vehicle in 178 wine samples from three *Vitis vinifera* cultivars; Strand B estimates the association between preferred alcoholic beverage and 10-year all-cause mortality in 1,629 adults of the NHANES I Epidemiologic Follow-up Study. Both strands execute on the shared containerised data-management platform described in Section 2.2.

2.2 Data management and execution platform

Both strands were served by a single containerised platform with five layers: a source layer holding the raw files, an extract-transform-load layer, a relational storage layer, an analytical workload layer, and an artefact layer that emits result tables, figures and a run manifest (Figure 2a). The container was provisioned with 2 virtual CPUs of an Intel Xeon processor at 2.10 GHz, 7 GiB of RAM and ephemeral local storage, which is representative of the smallest allocation a research group would request from a shared cloud or campus cluster.

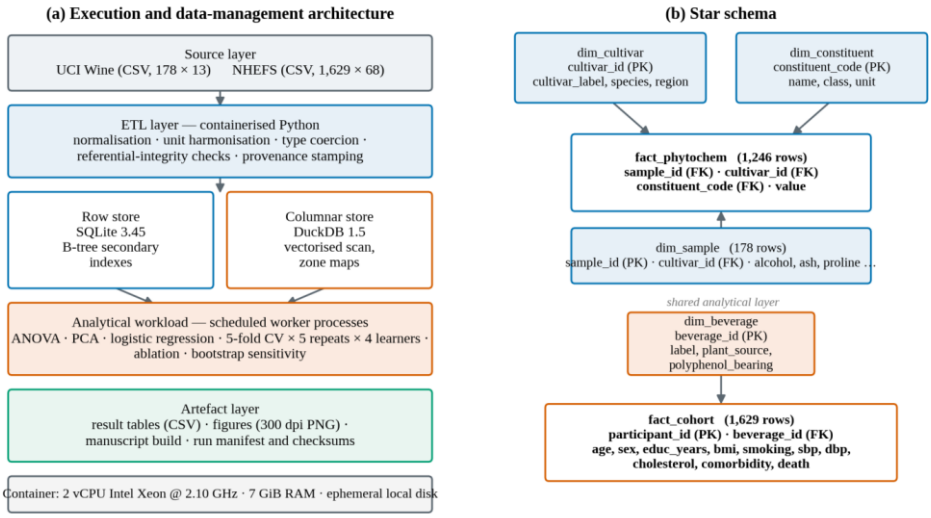

Figure 2. (a) Five-layer execution and data-management architecture, from raw files through extract-transform-load and relational storage to the scheduled analytical workload and emitted artefacts. (b) Star schema: two fact tables carry the observations and four dimension tables the descriptive attributes. The measurement fact table is held in long form so that new constituents can be added without schema migration.

Storage used a six-table star schema (Figure 2b). Two fact tables carry the observations: fact_phytochem holds one row per wine sample and constituent, and fact_cohort holds one row per cohort participant. Four dimension tables carry the descriptive attributes: dim_cultivar, dim_constituent, dim_sample and dim_beverage. The measurement fact table was deliberately held in long form, one row per sample and constituent rather than one column per constituent, so that new constituents or new assay panels can be added without altering the schema. dim_beverage carries an explicit plant_source attribute, which is the join point at which a botanical source can later be linked to a compositional reference such as Phenol-Explorer. Extraction applied unit harmonisation, type coercion, derivation of body-mass index, referential-integrity checks against every foreign key, and provenance stamping of the source file and load timestamp.

The same schema was materialised in two engines with different physical designs: SQLite 3.45, a row store with B-tree secondary indexes, and DuckDB 1.5, a columnar store with vectorised scans and zone maps (Raasveldt and Mühleisen, 2019). Four secondary indexes were defined on the fact-table foreign keys and on the outcome column. Five queries representative of the analysis were benchmarked: a grouped aggregate over the phytochemical facts (Q1), a selective filtered scan (Q2), a star join to the beverage dimension (Q3), a stratified cross-tabulation (Q4), and a partitioned window ranking (Q5). Each query was executed once to warm the cache and then fifteen times, and the median latency is reported. Data-volume scaling was assessed by replicating fact_cohort by factors of 1, 10, 100 and 1,000, up to 1.63 million rows, and re-timing Q3. Parallel behaviour was assessed by executing the cross-validation workload over 1, 2 and 4 worker processes on the 2-vCPU allocation, taking the fastest of three runs at each degree. Wall-clock time, cumulative CPU time and peak resident set size were recorded for each pipeline stage.

2.3 Phytochemical data set

Strand A used the Wine data set of the UCI Machine Learning Repository (Aeberhard and Forina, 1992), originally generated by Forina and colleagues and widely used as a reference chemometric benchmark. It reports a chemical analysis of 178 wine samples derived from three *Vitis vinifera* cultivars grown in the same region of Italy, with 13 constituents quantified per sample and no missing values. Seven variables relevant to polyphenol content and phenolic extraction were retained: total phenols, flavanoids, non-flavanoid phenols, proanthocyanins, colour intensity, hue and the OD280/OD315 ratio of diluted wines. Constituent values are reported in the original arbitrary analytical units; because all inference is between-group and scale-invariant, absolute calibration does not affect the conclusions.

2.4 Cohort data set

Strand B used the analytic file of the NHANES I Epidemiologic Follow-up Study (NHEFS), a prospective cohort assembled by the United States National Center for Health Statistics from participants examined in the first National Health and Nutrition Examination Survey during 1971–1975 and re-interviewed in 1982 (Madans et al., 1986). The publicly distributed analytic subset (Hernán and Robins, 2020) comprises 1,629 adults aged 25–74 years who smoked cigarettes at baseline and who had complete information at both visits. Baseline measurements include a physician-administered questionnaire, anthropometry, blood pressure and serum total cholesterol.

2.5 Exposure, outcome and covariates

The exposure was the participant's self-reported preferred alcoholic beverage at baseline, recorded in four mutually exclusive categories: beer, wine, spirits, or no stated preference. The primary contrast was wine preference versus all other categories combined; a secondary analysis compared all four categories with beer as the reference. The outcome was all-cause death recorded by the 1982 follow-up. Covariates were selected a priori as established determinants of both drinking pattern and mortality: age, sex, ethnicity, years of schooling, cigarettes per day, years of smoking, body-mass index, systolic blood pressure, serum total cholesterol, prevalent hypertension, prevalent diabetes, prevalent tumour, physical inactivity, drinking frequency and drinks per drinking occasion.

2.6 Statistical analysis

Strand A. Constituent concentrations were compared across cultivars by one-way analysis of variance with the Kruskal-Wallis test as a distribution-free confirmation, and the proportion of variance attributable to cultivar was expressed as eta squared. Pairwise contrasts used Tukey's honestly significant difference procedure at a family-wise error rate of 0.05. Multivariate structure was examined by principal component analysis on standardised variables and by Spearman rank correlation.

Strand B. Baseline characteristics were compared across beverage groups by one-way analysis of variance for continuous variables and the chi-squared test for categorical variables. Five nested logistic regression models were fitted, adding covariate blocks sequentially, so that the attenuation of the exposure coefficient could be attributed to specific confounder domains. Fewer than 5% of values were missing for systolic and diastolic blood pressure, serum cholesterol, drinks per occasion and schooling; these were imputed once at the sex-specific median, and a complete-case analysis was retained as a sensitivity check.

Predictive performance was assessed for four classifiers, namely unpenalised logistic regression, elastic-net logistic regression, random forest and gradient boosting, using stratified five-fold cross-validation repeated five times with fixed seeds; discrimination was summarised by the area under the receiver operating characteristic curve (AUC) and calibration by the Brier score. Feature ablation removed pre-specified covariate blocks and re-estimated performance under the identical resampling scheme. The robustness of the exposure estimate to unmeasured confounding was quantified with the E-value, the minimum strength of association that an unmeasured confounder would need with both the exposure and the outcome to explain away the observed estimate (VanderWeele and Ding, 2017). Pre-specified sensitivity analyses comprised complete-case analysis, exclusion of participants with prevalent tumour or diabetes, restriction to regular drinkers, and stratification by age and sex. Analyses used Python 3.11 with pandas, statsmodels, scipy and scikit-learn (Pedregosa et al., 2011); two-sided p values below 0.05 were treated as evidence against the null hypothesis.

3. RESULTS AND DISCUSSION

3.1 Platform behaviour: schema, storage engine and workload

The integrated schema comprised 3,067 rows across six tables (Table 1). Loading it took 19 ms in the row store and 37 ms in the columnar store, corresponding to throughputs of 164,000 and 83,000 rows per second respectively, and the resulting files occupied 156 KiB and 1,804 KiB (Table 2). The columnar engine's eleven-fold larger footprint at this scale reflects a fixed block and metadata overhead that is amortised only over much larger data; the entire integrated corpus for this study fits comfortably in the memory of a laptop, let alone a cloud node.

Table 1. Integrated star schema. Two fact tables carry the observations and four dimension tables the descriptive attributes shared by both analytical strands.

Table	Role	Rows	Columns	Grain
dim_cultivar	dimension	3	4	one row per cultivar
dim_constituent	dimension	7	4	one row per constituent
dim_sample	dimension	178	8	one row per wine sample
fact_phytochem	fact	1,246	4	one row per sample × constituent
dim_beverage	dimension	4	4	one row per beverage category
fact_cohort	fact	1,629	19	one row per cohort participant

Table 2. Extract-transform-load throughput and on-disk footprint of the integrated schema in a row store and a columnar store, measured on a 2-vCPU container.

Engine	Tables	Rows loaded	Load time (s)	Throughput (rows s^{-1})	On-disk size (KiB)
SQLite	6	3,067	0.019	163,996	156.0
DuckDB	6	3,067	0.037	83,363	1,804.0

Query latency depended on both the engine and the query shape (Table 3, Figure 3). On the native data volume the row store was faster for four of the five queries, most markedly for the selective filtered scan Q2 at 0.029 ms against 0.459 ms, because a B-tree index on a small table resolves a point predicate in a few page reads whereas a vectorised engine pays a fixed per-batch cost. Secondary indexes benefited the row store substantially, reducing Q1 by a factor of 2.5 and Q2 by a factor of 2.2, but changed the columnar engine by at most 8%, since its zone maps already prune blocks without an explicit index. The one query on which the columnar engine won at native scale was the partitioned window ranking Q5, at 1.89 ms against 2.86 ms, where vectorised execution over a whole column dominates row-at-a-time sorting.

Table 3. Median latency in milliseconds of five representative analytical queries by storage engine and index state, at the native data volume. Medians of fifteen executions after one warm-up execution.

Query	SQLite (no index)	SQLite (indexed)	DuckDB (no index)	DuckDB (indexed)
Q1 group aggregate	1.153	0.457	2.018	1.872
Q2 filtered scan	0.063	0.029	0.492	0.459
Q3 star join	0.726	0.694	1.635	1.641
Q4 stratified crosstab	1.092	1.029	1.535	1.520
Q5 window ranking	2.755	2.863	1.927	1.893

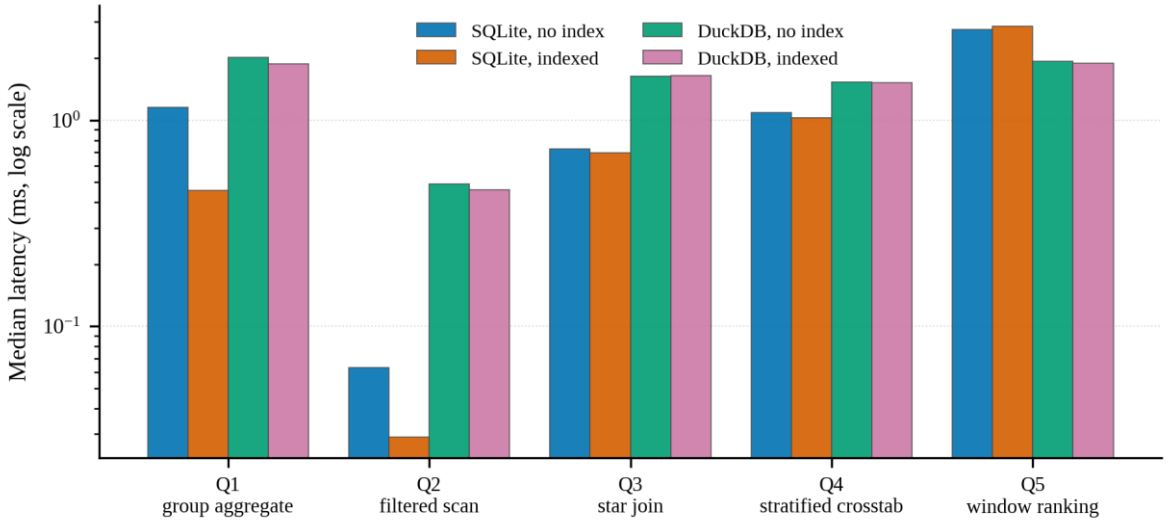

Figure 3. Median latency of five representative analytical queries in a row store (SQLite) and a columnar store (DuckDB), with and without secondary indexes, at the native data volume. The vertical axis is logarithmic. Medians of fifteen executions after one warm-up execution.

The ordering reverses with volume (Figure 4a). Replicating the cohort fact table to 16,290 rows already placed the two engines level, and by 1.63 million rows the star join took 1,077 ms in the row store against 32 ms in the columnar store, a 34-fold difference. Row-store latency grew slightly super-linearly, by a factor of 1,561 for a 1,000-fold increase in rows, whereas columnar latency grew only 14-fold over the same range. The practical reading is that engine choice should follow the anticipated federation scale rather than the current one: a single cohort is comfortably served by an embedded row store, but any design that anticipates pooling several cohorts, or joining against a full compositional reference database, should adopt a columnar engine from the outset, because migrating a schema after the analysis code has been written is considerably more expensive than choosing correctly at the start.

NaturalResourcesfor
Human Health
eISSN: 2583-1194

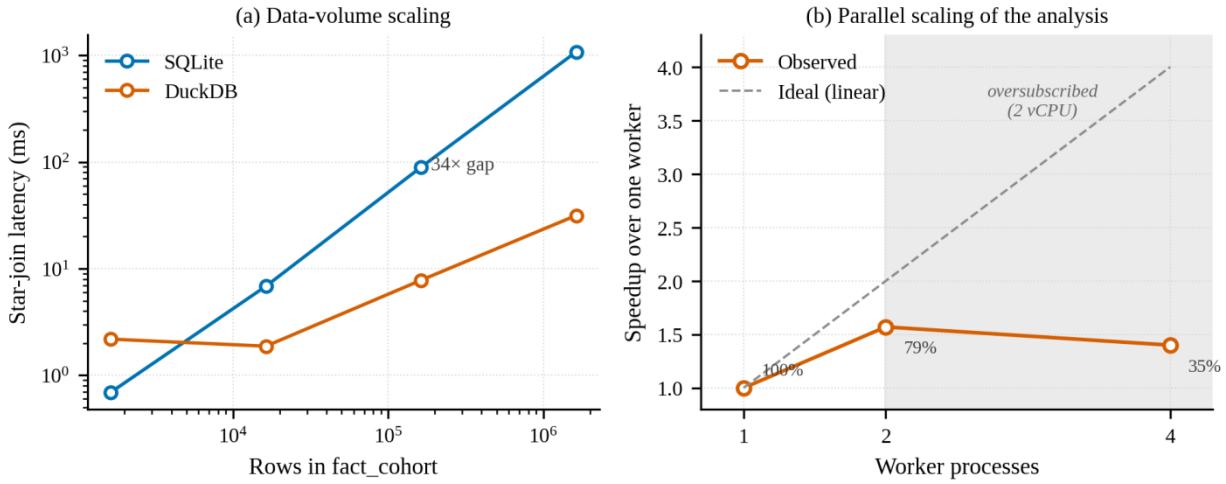

Figure 4. (a) Latency of the star join as the cohort fact table is replicated from 1,629 to 1,629,000 rows; both axes logarithmic. The engines cross over at roughly 16,000 rows and differ 34-fold at the largest volume. (b) Speedup of the cross-validation workload over one, two and four worker processes on a 2-vCPU allocation, with parallel efficiency annotated; the shaded region is oversubscribed.

The analytical workload was dominated by Strand B, which required 51.7 s of wall-clock time against 3.4 s for Strand A, with peak resident memory of 270 MiB and 215 MiB respectively (Table 4). Mean CPU utilisation was 49% and 48% of the two-vCPU allocation, that is, almost exactly one core: both stages are single-threaded by default, so half of the provisioned capacity was idle throughout. Distributing the cross-validation folds over two worker processes reduced wall-clock time from 23.8 s to 15.1 s, a speedup of 1.57 and a parallel efficiency of 79%, the shortfall from linear being attributable to process start-up, data serialisation to the workers and contention for shared memory bandwidth. Requesting four workers on the same two cores was actively counterproductive, lengthening the run to 17.0 s and cutting efficiency to 35% (Figure 4b).

Table 4. Execution profile of the analytical workload. Stage rows report wall-clock time, cumulative CPU time, mean utilisation of the two-vCPU allocation and peak resident memory; worker rows report speedup of the cross-validation workload.

Component	Wall-clock (s)	CPU time (s)	CPU utilisation (%)	Peak RSS (MiB)
Strand A: phytochemical analysis	3.38	3.24	47.9	215.4
Strand B: cohort analysis	51.71	50.88	49.2	270.3
Cross-validation, 1 worker	23.78	—	speedup 1.00	efficiency 100%
Cross-validation, 2 workers	15.13	—	speedup 1.57	efficiency 79%
Cross-validation, 4 workers	16.97	—	speedup 1.40	efficiency 35%

That oversubscription penalty is worth stating plainly, because it is the same phenomenon that governs throughput in shared research clusters. Capacity added beyond the point of effective concurrency does not merely fail to help; it costs measurable throughput, and a scheduler that packs jobs by requested worker count rather than by measured core occupancy will systematically over-commit. For statistical workloads of this shape, the operationally useful summary is that the correct allocation is one worker per physical core, and that the return on further cores is negative unless the code is restructured.

3.2 Cultivar accounts for a large share of phenolic variance

All seven constituents differed significantly across the three cultivars (Table 5; all $p < 0.001$ by both analysis of variance and the Kruskal-Wallis test). The magnitude of the differences, however, is more informative than their statistical significance. Cultivar accounted for 72.8% of the variance in flavanoid content, 68.5% in the OD280/OD315 ratio, 58.0% in colour intensity and 51.7% in total phenols, but only 24.0% in non-flavanoid phenols and 25.7% in proanthocyanins.

Table 5. Phenolic and colour constituents of 178 wine samples from three *Vitis vinifera* cultivars grown in the same Italian region. Values are mean ± standard deviation in the original arbitrary analytical units. F and p from one-way analysis of variance; η^2 is the proportion of variance attributable to cultivar.

Constituent	Cultivar 1	Cultivar 2	Cultivar 3	F	η^2	p
Total phenols	2.84 ± 0.34	2.26 ± 0.55	1.68 ± 0.36	93.7	0.517	<0.001
Flavanoids	2.98 ± 0.40	2.08 ± 0.71	0.78 ± 0.29	233.9	0.728	<0.001
Non-flavanoid phenols	0.29 ± 0.07	0.36 ± 0.12	0.45 ± 0.12	27.6	0.240	<0.001
Proanthocyanins	1.90 ± 0.41	1.63 ± 0.60	1.15 ± 0.41	30.3	0.257	<0.001
Colour intensity	5.53 ± 1.24	3.09 ± 0.92	7.40 ± 2.31	120.7	0.580	<0.001
Hue	1.06 ± 0.12	1.06 ± 0.20	0.68 ± 0.11	101.3	0.537	<0.001
OD280/OD315	3.16 ± 0.36	2.79 ± 0.50	1.68 ± 0.27	190.0	0.685	<0.001

Mean flavanoid content was 2.98 ± 0.40 units in Cultivar 1 and 0.78 ± 0.29 units in Cultivar 3, a 3.8-fold difference; total phenols differed 1.7-fold between the same two groups (Figure 5). Tukey contrasts separated all three cultivars for flavanoids and for total phenols. Non-flavanoid phenols ran in the opposite direction, being highest in the cultivar poorest in flavanoids (0.45 ± 0.12 versus 0.29 ± 0.07 units), so that total phenolic content understates how differently the phenolic fraction is composed.

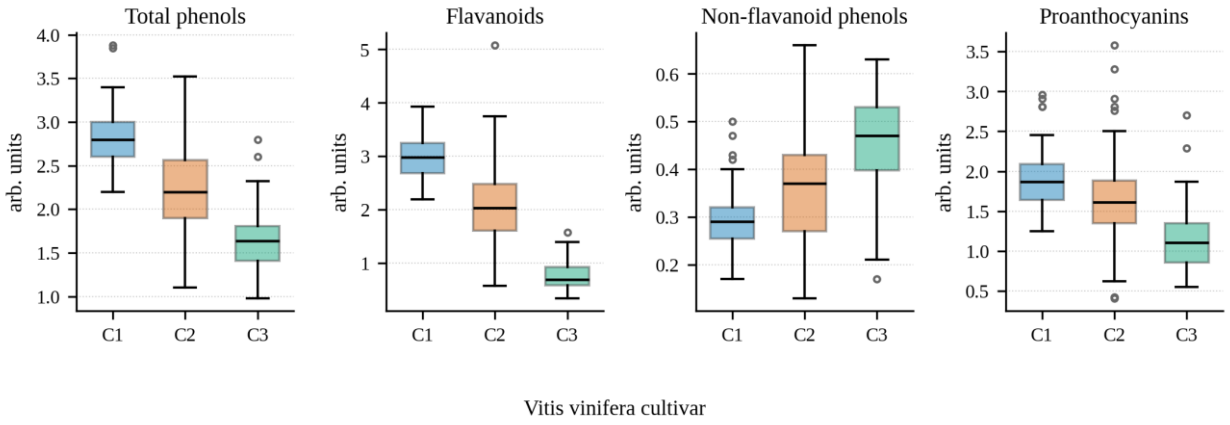

Figure 5. Distribution of four phenolic constituents across the three *Vitis vinifera* cultivars (C1-C3). Boxes show the interquartile range and median; whiskers extend to 1.5 times the interquartile range. Mean flavanoid content differs 3.8-fold between C1 and C3, whereas non-flavanoid phenols run in the opposite direction.

This is the first component of the argument. Samples grown in one region, analysed in one laboratory and sold under one product name differ several-fold in the constituents most often invoked as the active principle. Any epidemiological variable that records only the presence or volume of the beverage carries that variation as irreducible exposure misclassification. Because the misclassification is non-differential with respect to the outcome, its expected effect is attenuation towards the null, which means that a genuinely protective compound could be obscured just as easily as a spurious association could be generated.

3.3 The phenolic profile is dominated by a single axis

Principal component analysis of the seven standardised constituents returned a first component explaining 55.3% of total variance and a second explaining 18.4%, cumulatively 73.8% (Figure 6). The first component loaded positively and strongly on flavanoids (0.472), OD280/OD315 (0.451), total phenols (0.434) and proanthocyanins (0.359), and negatively on non-

flavanoid phenols (-0.321). It therefore represents a general 'flavanoid richness' axis along which the three cultivars separate almost completely.

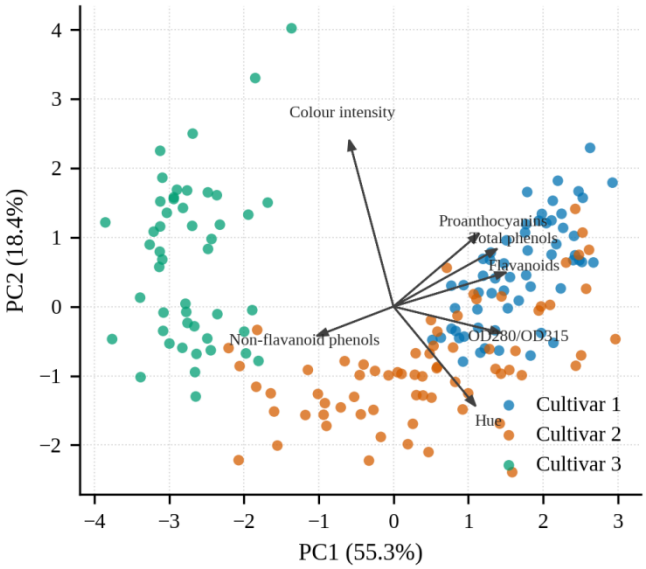

Figure 6. Principal component biplot of seven standardised phenolic and colour constituents. The first two components explain 73.8% of total variance. The first component is a general flavanoid-richness axis on which the three cultivars separate almost completely.

Spearman correlations were correspondingly high within the flavanoid block, reaching $\varrho = 0.88$ between total phenols and flavanoids, and negative between that block and non-flavanoid phenols (Figure 7). Two consequences follow for exposure assessment. First, because the constituents are so strongly collinear, a single well-chosen analytical marker would capture most of the compositional information; measuring one polyphenol well is more informative than recording beverage volume. Second, the dominant axis is a cultivar axis, and cultivar is not recorded in any dietary questionnaire in routine epidemiological use, nor is there a field in which to store it.

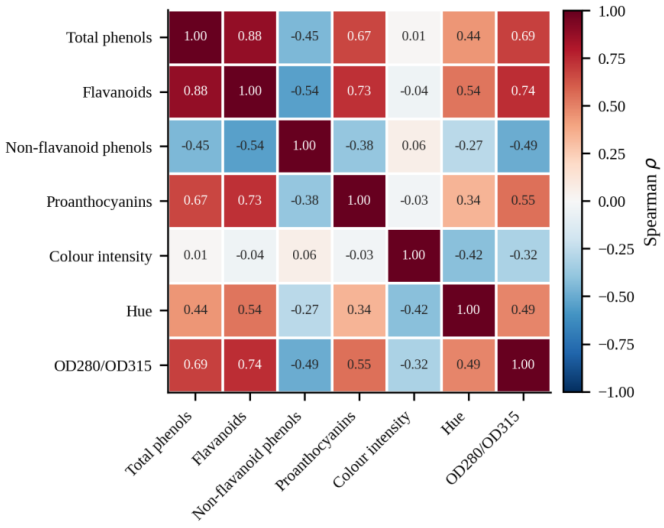

Figure 7. Spearman rank correlation matrix of the seven constituents. Strong positive correlation within the flavanoid block ($\varrho = 0.88$ between total phenols and flavanoids) indicates that a single well-chosen analytical marker would capture most of the compositional information.

3.4 Wine drinkers differ systematically from other drinkers

The cohort comprised 1,629 adults, of whom 318 (19.5%) died by the 1982 follow-up. Preferred beverage was beer in 569 participants, wine in 132, spirits in 512 and unstated in 416. Baseline characteristics differed sharply across these groups (Table 6).

Table 6. Baseline characteristics of 1,629 NHEFS participants by preferred alcoholic beverage. Continuous variables are mean ± standard deviation and compared by one-way analysis of variance; categorical variables are n (%) and compared by the chi-squared test.

Characteristic	Beer (n = 569)	Wine (n = 132)	Spirits (n = 512)	No preference (n = 416)	p
Age at baseline (years)	41.2 ± 12.1	40.0 ± 11.4	45.1 ± 12.1	47.4 ± 11.4	<0.001
Body-mass index (kg m⁻²)	24.7 ± 4.5	23.4 ± 4.3	25.0 ± 4.5	25.4 ± 5.9	<0.001
Systolic blood pressure (mmHg)	128.8 ± 17.5	124.3 ± 16.5	128.5 ± 18.8	129.9 ± 20.2	0.026
Diastolic blood pressure (mmHg)	79.6 ± 10.2	76.9 ± 10.7	76.9 ± 10.2	76.5 ± 10.4	<0.001
Serum total cholesterol (mg dL⁻¹)	211.2 ± 40.2	218.7 ± 46.8	228.3 ± 47.5	222.0 ± 46.3	<0.001
Cigarettes per day	22.1 ± 12.6	19.7 ± 13.1	20.3 ± 11.1	18.9 ± 10.8	<0.001
Years of smoking	23.0 ± 12.3	20.8 ± 11.0	25.7 ± 12.1	27.6 ± 11.7	<0.001
Drinks per drinking occasion	3.7 ± 3.4	2.1 ± 1.4	3.1 ± 2.6	2.3 ± 0.5	<0.001
Years of schooling	10.9 ± 3.1	12.6 ± 2.9	12.0 ± 2.6	10.0 ± 3.2	<0.001
Female, n (%)	175 (30.8)	82 (62.1)	299 (58.4)	274 (65.9)	<0.001
Non-white, n (%)	87 (15.3)	15 (11.4)	61 (11.9)	52 (12.5)	0.324
Prevalent hypertension, n (%)	41 (7.2)	4 (3.0)	38 (7.4)	47 (11.3)	0.010
Prevalent diabetes, n (%)	1 (0.2)	0 (0.0)	5 (1.0)	8 (1.9)	0.020
Little or no exercise, n (%)	220 (38.7)	34 (25.8)	183 (35.7)	198 (47.6)	<0.001
Prevalent tumour, n (%)	8 (1.4)	2 (1.5)	13 (2.5)	15 (3.6)	0.134
All-cause death by 1982, n (%)	104 (18.3)	13 (9.8)	92 (18.0)	109 (26.2)	<0.001

Participants who preferred wine had the most years of schooling (12.6 ± 2.9 years versus 10.9 ± 3.1 for beer and 10.0 ± 3.2 for those with no preference), the lowest body-mass index (23.4 ± 4.3 kg m⁻²), the lowest systolic blood pressure (124.3 ± 16.5 mmHg), the lowest prevalence of hypertension (3.0% versus 11.3%), the lowest prevalence of physical inactivity (25.8%

versus 47.6%) and the fewest drinks per drinking occasion (2.1 ± 1.4 versus 3.7 ± 3.4 for beer). They were also the youngest group and, at 62.1% female, the second most female. Every one of these differences favours survival independently of anything the beverage contains.

Crude 10-year mortality tracked this gradient closely: 9.8% among wine drinkers, 18.0% among spirits drinkers, 18.3% among beer drinkers and 26.2% among those with no stated preference ($p < 0.001$; Figure 8). The rank order of mortality across the four groups is the rank order of their baseline socioeconomic and behavioural advantage, not of the polyphenol content of the beverages: spirits, which contain essentially no polyphenols, and beer, which contains modest amounts, gave almost identical crude risks.

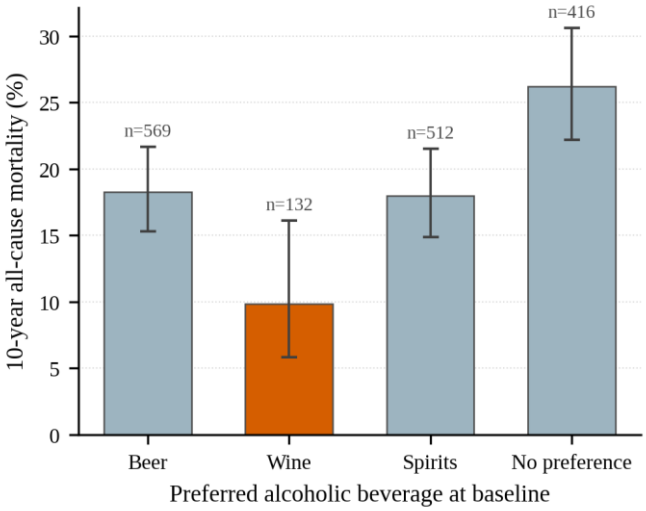

Figure 8. Crude 10-year all-cause mortality by preferred alcoholic beverage at baseline, with Wilson 95% confidence intervals. Group sizes are shown above each bar. The rank order of mortality follows the rank order of baseline socioeconomic and behavioural advantage rather than the polyphenol content of the beverages.

3.5 The apparent benefit of wine attenuates under adjustment

In the unadjusted model, wine preference was associated with a 57% lower odds of death (odds ratio 0.43, 95% CI 0.24–0.77; $p = 0.004$). Adding age, sex and ethnicity moved the estimate to 0.61 (0.32–1.18); adding schooling, smoking intensity and duration, and body-mass index moved it to 0.68 (0.35–1.33); adding clinical covariates moved it to 0.70 (0.35–1.38); and adding drinking frequency and quantity left it essentially unchanged at 0.70 (0.35–1.39; $p = 0.307$). Fifty-eight per cent of the crude log-odds association was removed by covariate adjustment, and the confidence interval crossed unity from the second model onwards (Table 7, Figure 9).

Table 7. Odds ratios for 10-year all-cause mortality comparing wine preference with all other beverage preferences under sequentially adjusted logistic regression models.

Model	Odds ratio (95% CI)	p	n
Model 1: unadjusted	0.43 (0.24–0.77)	0.004	1629
Model 2: + age, sex, ethnicity	0.61 (0.32–1.18)	0.145	1629
Model 3: + schooling, smoking, BMI	0.68 (0.35–1.33)	0.263	1629
Model 4: + clinical covariates	0.70 (0.35–1.38)	0.304	1629
Model 5: + drinking frequency and quantity	0.70 (0.35–1.39)	0.307	1629

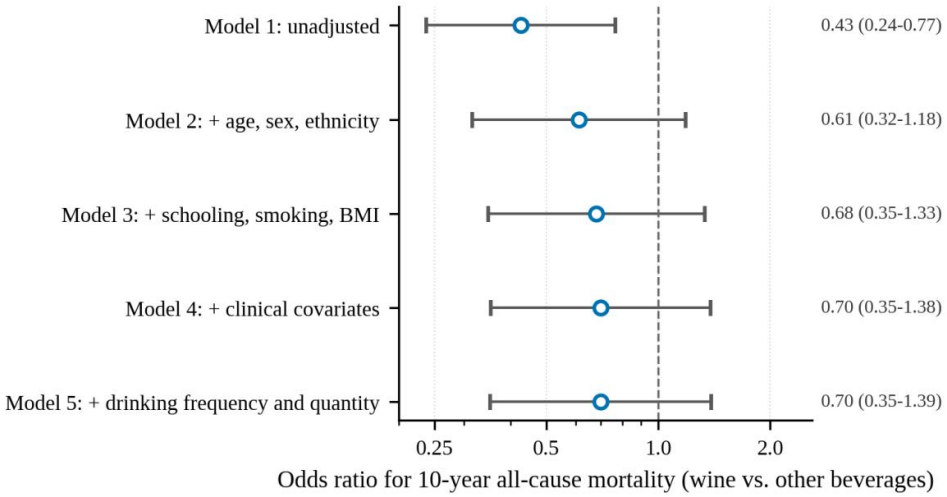

Figure 9. Odds ratios and 95% confidence intervals for 10-year all-cause mortality comparing wine preference with all other beverage preferences, under sequentially added covariate blocks (Models 1–5). The horizontal axis is on a logarithmic scale; the dashed line marks the null.

The stepwise pattern is informative. The single largest attenuation came from demographic adjustment, and the second largest from schooling and smoking; the clinical block and the drinking-behaviour block added almost nothing. In other words, the crude association was generated primarily by who chooses wine rather than by how much or how often they drink. In the four-category comparison with beer as reference (Table 8), the adjusted odds ratio for wine was 0.66 (0.32–1.38) and for spirits 0.75 (0.50–1.12), while participants with no stated beverage preference had a higher adjusted odds of death (1.62, 0.93–2.82). That the largest adjusted signal in the data belongs to the group defined by the absence of a preference rather than by any polyphenol-bearing beverage is difficult to reconcile with a compound-specific mechanism.

Table 8. Adjusted odds ratios for 10-year all-cause mortality by beverage preference category, with beer as the reference group. Adjusted for all covariates in Model 5.

Beverage preference	Adjusted odds ratio (95% CI)	p
Beer (reference)	1.00 (reference)	—
No preference	1.62 (0.93–2.82)	0.089
Spirits	0.75 (0.50–1.12)	0.155
Wine	0.66 (0.32–1.38)	0.271

In the fully adjusted model the covariates behaved as expected for this cohort: each additional year of age raised the odds of death by 8.8% (1.088, 1.062–1.114), female sex was protective (0.66, 0.47–0.92), each additional year of schooling was associated with 5% lower odds (0.95, 0.90–1.00; p = 0.033), each additional year of smoking raised the odds by 2.9% (1.029, 1.006–1.052), and prevalent hypertension raised them by 85% (1.85, 1.18–2.92). The internal coherence of these estimates supports the adequacy of the model specification.

3.6 Beverage type carries no independent predictive information

Association and prediction answer different questions, and a variable can be a valid causal factor without improving prediction. Nevertheless, if beverage preference indexed a meaningful polyphenol exposure, some predictive contribution would be expected. All four classifiers discriminated 10-year mortality well and almost identically: AUC 0.832 for both unpenalised and elastic-net logistic regression, 0.825 for random forest and 0.814 for gradient boosting, with Brier scores between 0.114 and 0.120 (Table 9, Figure 10). The absence of any advantage for the flexible learners indicates that the

outcome surface is close to additive in these covariates. Decile calibration of the logistic model was acceptable across the range of predicted risk (Figure 11).

Table 9. Discrimination and calibration of four classifiers predicting 10-year all-cause mortality. Stratified five-fold cross-validation repeated five times; the interval is the mean ± 1.96 standard deviations across repeats.

Classifier	AUC (95% CI)	Brier score
Logistic regression	0.832 (0.827–0.837)	0.1144
Elastic-net logistic	0.832 (0.827–0.837)	0.1142
Random forest	0.825 (0.823–0.827)	0.1159
Gradient boosting	0.814 (0.805–0.822)	0.1198

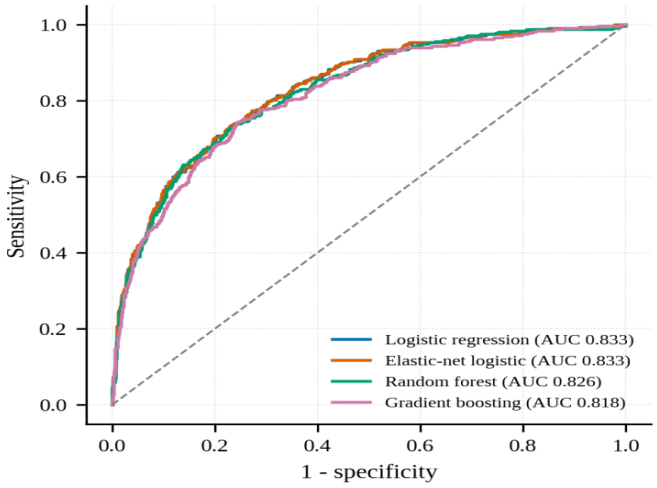

Figure 10. Receiver operating characteristic curves for four classifiers predicting 10-year all-cause mortality, based on out-of-fold predictions from stratified five-fold cross-validation repeated five times. Discrimination is essentially identical across model families.

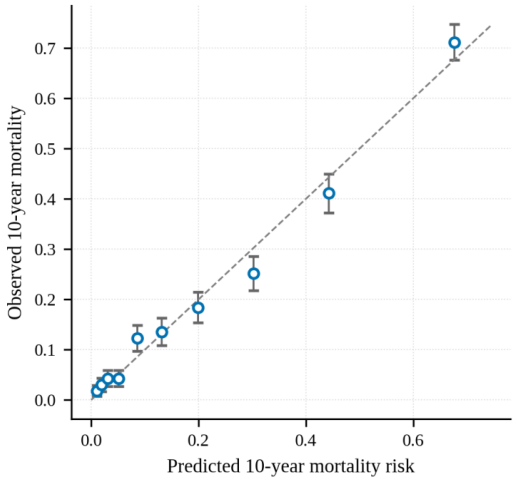

Figure 11. Calibration of the cross-validated logistic regression model by decile of predicted risk. Error bars are binomial standard errors; the dashed line indicates perfect calibration.

Feature ablation was unambiguous (Table 10). Removing the beverage-type indicator changed discrimination by +0.0003 AUC, and removing all three alcohol-related variables changed it by +0.0016 AUC; both differences are within resampling noise and are, if anything, in the direction of slight improvement. By contrast, removing the clinical measurement block cost 0.0045 AUC and removing the smoking variables cost 0.0008 AUC. Beverage preference thus adds no information that is not already contained in age, sex, schooling, smoking and clinical status, which is precisely what would be expected if it were a marker of those factors rather than an exposure in its own right.

Table 10. Feature ablation. Change in cross-validated discrimination when pre-specified covariate blocks are removed from the logistic regression model.

Feature set	Features (n)	AUC	ΔAUC	Brier score
Full model (all covariates)	17	0.832	+0.0000	0.1144
Without beverage-type indicator	16	0.832	+0.0003	0.1142
Without drinking frequency and quantity	15	0.833	+0.0013	0.1139
Without all alcohol-related variables	14	0.833	+0.0016	0.1137
Without smoking variables	15	0.831	-0.0008	0.1149
Without clinical measurements	11	0.827	-0.0045	0.1160

3.7 Sensitivity to unmeasured confounding

The fully adjusted odds ratio of 0.70 corresponds to an approximate risk ratio of 0.84 and an E-value of 1.68 (Figure 12). An unmeasured confounder would need to be associated with both wine preference and mortality by a risk ratio of at least 1.68, above and beyond every measured covariate, to reduce the estimate to the null. Associations of that strength are entirely plausible in this setting: household income, occupational class, dietary quality and access to health care were not measured here, and each is known to relate to both beverage choice and survival at least this strongly. Because the confidence interval already includes the null, the corresponding E-value for the interval limit is 1.00.

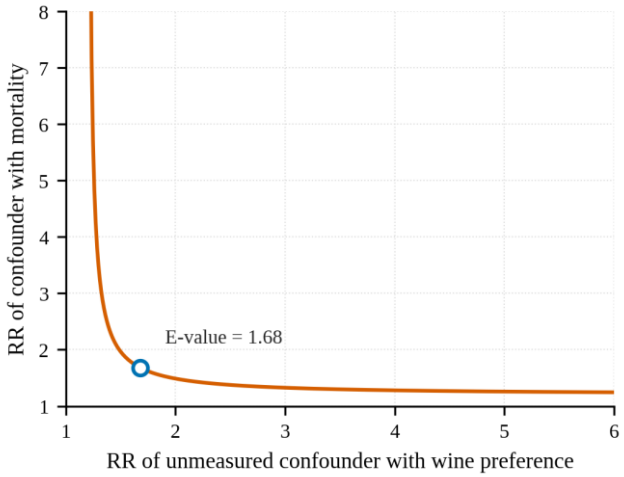

Figure 12. Bias analysis for the fully adjusted estimate. The curve shows combinations of confounder-exposure and confounder-outcome associations that would be required to explain away the observed association; the marked point is the E-value of 1.68.

The point estimate was stable across all pre-specified sensitivity analyses (Table 11), ranging from 0.63 when participants with prevalent tumour or diabetes were excluded to 0.85 among men, with every confidence interval including unity. Stratified estimates were more protective in women (0.53, 0.17–1.64) than in men (0.85, 0.34–2.17) and in participants under 50 years

(0.40, 0.09–1.71) than in those aged 50 years or more (0.79, 0.34–1.86), but the strata were small and the intervals overlapped almost completely, so these contrasts should not be interpreted as effect modification.

Table 11. Sensitivity analyses for the fully adjusted association between wine preference and 10-year all-cause mortality.

Analysis	n	Odds ratio (95% CI)	p
Primary (Model 5, full analytic sample)	1629	0.70 (0.35–1.39)	0.307
Complete-case analysis (no imputation)	1141	0.66 (0.31–1.40)	0.283
Excluding prevalent tumour or diabetes	1577	0.63 (0.31–1.28)	0.198
Restricted to regular drinkers	1417	0.73 (0.36–1.46)	0.374
Aged < 50 years	1064	0.40 (0.09–1.71)	0.216
Aged ≥ 50 years	565	0.79 (0.34–1.86)	0.593
Men	799	0.85 (0.34–2.17)	0.738
Women	830	0.53 (0.17–1.64)	0.268

3.8 Joint interpretation

Read together, the two strands describe a proxy that fails at both ends. Upstream, the vehicle is not compositionally stable: cultivar alone, with region held constant, produces a 3.8-fold spread in flavanoid content and reverses the direction of the non-flavanoid fraction. Downstream, the behaviour of consuming the vehicle is strongly confounded: more than half of the crude association with mortality is removed by routinely measured covariates, and the residue is fragile to unmeasured confounding of unremarkable strength. A variable that is noisy as a measure of dose and contaminated as a measure of behaviour cannot support inference about a specific compound.

This reading is consistent with the direction the wider literature has taken. Genetic evidence in over 500,000 Chinese adults found no cardioprotective effect of alcohol once genotype was used as an instrument (Millwood et al., 2019), and a meta-analysis of 107 cohort studies found that the apparent benefit of low-volume drinking disappeared in studies that avoided abstainer-reference bias (Zhao et al., 2023). Our results add a complementary, chemical explanation for why the wine-specific version of the hypothesis proved difficult to replicate. It should be emphasised that these findings do not weigh against a role for dietary polyphenols in health. Cohorts that estimate flavonoid intake from compositional databases rather than from a single carrier beverage continue to report robust inverse associations with mortality (Bondonno et al., 2019; Grosso et al., 2017), and recent work indicates that the diversity of flavonoid sources, not merely the total, predicts risk (Parmenter et al., 2025). The distinction is one of measurement: those studies estimate a compound, ours interrogates a vehicle.

The practical implication for natural-product research is that exposure should be quantified at the level of the compound wherever possible, and the benchmark results indicate that the computational obstacle to doing so is negligible. Three approaches are available. Food composition databases such as Phenol-Explorer allow intake to be estimated from dietary records with explicit uncertainty ranges (Neveu et al., 2010); in the schema presented here this is a join from dim_beverage or a food dimension onto a compositional reference table, which the columnar engine resolves in tens of milliseconds even at a million fact rows. Analytical characterisation of the specific product consumed, as performed in Strand A, removes cultivar-driven variance and is naturally accommodated by the long-form measurement fact table, which absorbs new constituents without schema migration. Circulating or urinary polyphenol metabolites provide an integrated biomarker of internal dose that is unaffected by reporting error, although interindividual differences in absorption and microbial metabolism must be accounted for (Del Rio et al., 2013; Manach et al., 2005). Where none of these is available, an observed association with a carrier food should be reported as an association with that food, not with the compound it contains.

3.9 Strengths and limitations

The principal strength of this work is its design: pairing a compositional analysis of the exposure vehicle with an epidemiological analysis of the same vehicle allows two failure modes of proxy exposures to be examined in one study, and running both on an instrumented, schema-governed platform makes every step auditable and reproducible. All data are openly available, the code is released, and the pre-specified ablation and quantitative bias analyses guard against selective reporting.

Several limitations qualify the findings. First, and most important, every participant in the NHEFS analytic file smoked cigarettes at baseline, which explains the high 10-year mortality of 19.5% and means the estimates should not be generalised to non-smoking populations. Second, only 132 participants preferred wine and only 13 of them died, so the analysis has limited power; the confidence interval excludes odds ratios below 0.35 and above 1.39, which rules out large effects in either direction but not modest ones. Third, exposure was a single self-reported preference at baseline, with no measure of volume of wine specifically, no repeat assessment and no information on wine style or cultivar. Fourth, the outcome was all-cause mortality; cause-specific analyses might reveal associations that the composite masks. Fifth, the baseline examination took place in 1971–1975, and both drinking patterns and the confounding structure of beverage choice have changed since. Sixth, the two strands derive from different populations and eras and are combined argumentatively rather than statistically; the compositional data quantify the variability that the cohort exposure ignores, but they are not the wines that the cohort drank. Seventh, the benchmarks were obtained on a single 2-vCPU container, so the absolute latencies are specific to that allocation and the parallel results characterise scaling only up to the point of oversubscription; multi-node behaviour, network-attached storage and concurrent multi-tenant load were not examined. Finally, the phenolic constituents are reported in arbitrary analytical units, so absolute intakes cannot be inferred from Strand A.

4. CONCLUSION

Beverage type is a weak instrument for studying plant polyphenols. Cultivar alone explains up to 73% of the variance in phenolic constituents of wines grown in a single region, producing a 3.8-fold spread in flavanoid content that no dietary questionnaire records and no conventional study schema can store. In a national cohort followed for 10 years, the substantial crude survival advantage of wine drinkers was 58% attenuated by adjustment for demographic, socioeconomic, behavioural and clinical covariates, was fragile to unmeasured confounding of modest strength (E-value 1.68), and contributed nothing to predictive discrimination. The apparent benefit is better explained by who drinks wine than by what wine contains. Research on natural resources for human health should therefore measure the bioactive rather than its vehicle, and the infrastructure results reported here show that the barrier to doing so is not computational: a normalised star schema over both compositional and cohort data loaded in tens of milliseconds, and a columnar engine answered federated queries 34 times faster than a row store at 1.63 million fact rows on a container costing a few cents an hour. Future work should extend the two-strand design to other carrier-based exposures such as tea, coffee, cocoa and culinary herbs; federate several cohorts and a full compositional reference database within the same schema; and characterise the resulting workload on a multi-node cluster, where the oversubscription penalty observed here at four workers on two cores becomes a scheduling problem rather than a local one.

Acknowledgements

The authors thank the United States National Center for Health Statistics for collecting and distributing the NHANES I Epidemiologic Follow-up Study, and the UCI Machine Learning Repository and the original donor for maintaining the Wine data set.

Author Contributions

Deepak Yadav: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Visualization, Writing—Original Draft.

Savita Kumari Sheoran: Conceptualization, Methodology, Validation, Supervision, Writing—Review & Editing.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest

The authors declare that there is no conflict of interest. Uniform disclosure forms have been completed for all authors.

Ethics Statement

Ethical approval was not required for this study. Both data sets are publicly available, fully de-identified secondary sources.

Data Availability Statement

All data supporting the findings of this study are publicly available. The Wine data set is available from the UCI Machine Learning Repository (https://doi.org/10.24432/C5PC7J) and is distributed with scikit-learn. The NHANES I Epidemiologic Follow-up Study analytic file is available from the National Center for Health Statistics and is redistributed in the causaldata package (https://doi.org/10.32614/CRAN.package.causaldata). The schema definition, extract-transform-load code, benchmark harness and analysis scripts that reproduce every table and figure in this manuscript are available from the corresponding author on request.

REFERENCES

[1] Abadi, D., Boncz, P., Harizopoulos, S., Idreos, S., Madden, S., 2013. The design and implementation of modern column-oriented database systems. Foundations and Trends in Databases 5(3), 197–280. doi:10.1561/1900000024.

[2] Aeberhard, S., Forina, M., 1992. Wine [Dataset]. UCI Machine Learning Repository. doi:10.24432/C5PC7J.

[3] Bondonno, N.P., Dalgaard, F., Kyrø, C., Murray, K., Bondonno, C.P., Lewis, J.R., Croft, K.D., Gislason, G., Scalbert, A., Cassidy, A., Tjønneland, A., Overvad, K., Hodgson, J.M., 2019. Flavonoid intake is associated with lower mortality in the Danish Diet Cancer and Health Cohort. Nature Communications 10, 3651. doi:10.1038/s41467-019-11622-x.

[4] Del Rio, D., Rodriguez-Mateos, A., Spencer, J.P.E., Tognolini, M., Borges, G., Crozier, A., 2013. Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases. Antioxidants & Redox Signaling 18(14), 1818–1892. doi:10.1089/ars.2012.4581.

[5] El Rayess, Y., Nehme, N., Azzi-Achkouty, S., Julien, S.G., 2024. Wine phenolic compounds: chemistry, functionality and health benefits. Antioxidants 13(11), 1312. doi:10.3390/antiox13111312.

[6] Grosso, G., Micek, A., Godos, J., Pajak, A., Sciacca, S., Galvano, F., Giovannucci, E.L., 2017. Dietary flavonoid and lignan intake and mortality in prospective cohort studies: systematic review and dose-response meta-analysis. American Journal of Epidemiology 185(12), 1304–1316. doi:10.1093/aje/kww207.

[7] Hernán, M.A., Robins, J.M., 2020. Causal Inference: What If. Chapman & Hall/CRC, Boca Raton, FL.

[8] Madans, J.H., Kleinman, J.C., Cox, C.S., Barbano, H.E., Feldman, J.J., Cohen, B., Finucane, F.F., Cornoni-Huntley, J., 1986. 10 years after NHANES I: report of initial followup, 1982–84. Public Health Reports 101(5), 465–473.

[9] Manach, C., Williamson, G., Morand, C., Scalbert, A., Rémésy, C., 2005. Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies. American Journal of Clinical Nutrition 81(1), 230S–242S. doi:10.1093/ajcn/81.1.230S.

[10] Millwood, I.Y., Walters, R.G., Mei, X.W., Guo, Y., Yang, L., Bian, Z., Bennett, D.A., Chen, Y., Dong, C., Hu, R., Zhou, G., Yu, B., Jia, W., Parish, S., Clarke, R., Davey Smith, G., Collins, R., Holmes, M.V., Li, L., Peto, R., Chen, Z., 2019. Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China. The Lancet 393(10183), 1831–1842. doi:10.1016/S0140-6736(18)31772-0.

[11] Neveu, V., Perez-Jiménez, J., Vos, F., Crespy, V., du Chaffaut, L., Mennen, L., Knox, C., Eisner, R., Cruz, J., Wishart, D., Scalbert, A., 2010. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. Database 2010, bap024. doi:10.1093/database/bap024.

[12] Parmenter, B.H., Thompson, A.S., Bondonno, N.P., Jennings, A., Murray, K., Perez-Cornago, A., Hodgson, J.M., Tresserra-Rimbau, A., Kühn, T., Cassidy, A., 2025. High diversity of dietary flavonoid intake is associated with a lower risk of all-cause mortality and major chronic diseases. Nature Food 6, 668–680. doi:10.1038/s43016-025-01176-1.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

[14] Raasveldt, M., Mühleisen, H., 2019. DuckDB: an embeddable analytical database. Pp. 1981–1984 in Proceedings of the 2019 International Conference on Management of Data. Amsterdam, Netherlands: ACM. doi:10.1145/3299869.3320212.
[15] Renaud, S., de Lorgeril, M., 1992. Wine, alcohol, platelets, and the French paradox for coronary heart disease. The Lancet 339(8808), 1523–1526. doi:10.1016/0140-6736(92)91277-F.
[16] VanderWeele, T.J., Ding, P., 2017. Sensitivity analysis in observational research: introducing the E-value. Annals of Internal Medicine 167(4), 268–274. doi:10.7326/M16-2607.
[17] Waterhouse, A.L., 2002. Wine phenolics. Annals of the New York Academy of Sciences 957(1), 21–36. doi:10.1111/j.1749-6632.2002.tb02903.x.
[18] Zhao, J., Stockwell, T., Naimi, T., Churchill, S., Clay, J., Sherk, A., 2023. Association between daily alcohol intake and risk of all-cause mortality: a systematic review and meta-analyses. JAMA Network Open 6(3), e236185. doi:10.1001/jamanetworkopen.2023.6185.