Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Long-Term Relationship Between Animal Product Consumption and Cancer Incidence: A Cointegration- and ARIMAX-Based Approach

Alessia Spada
alessia.spada@unifg.it

University of Foggia
Michele Tomaiuolo
Agorà Biomedical Sciences
Elisa Pia Amorusi
Agorà Biomedical Sciences
Nicholas Calà
Agorà Biomedical Sciences
Raffaele Ianzano
Agorà Biomedical Sciences
Pasquale Ieluzzi
Agorà Biomedical Sciences
Giovanni Emanuele Ricciardi
Vita-Salute San Raffaele University
Antonio Tucci
Agorà Biomedical Sciences

Article

Keywords: cointegration, ARIMAX model, Dairy, Meat, Breast Cancer, Prostate Cancer

Posted Date: September 19th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7489565/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on March 13th, 2026. See the published version at https://doi.org/10.1038/s41598-026-42068-z.

Abstract

Understanding how dietary habits shape the long-term incidence of hormone-sensitive cancers remains a major challenge. Conventional approaches, often constrained by short follow-up periods or static methods, risk producing spurious associations and weak conclusions. In this study, we analyzed exceptionally comprehensive Italian national time series (1961–2020 for meat and dairy consumption; 1984–2020 for cancer incidence) to investigate the association between diet and the development of breast and prostate cancer.

We first employed Principal Component Analysis (PCA) to synthesize consumption data into a single index (PC1), thereby reducing multicollinearity among variables. We then applied a rigorous econometric framework that combined with ARIMAX modeling, designed to distinguish genuine long-term dynamics from superficial statistical associations.

The analyses revealed evidence of cointegration between consumption and cancer incidence for both malignancies. In breast cancer, the optimal ARIMAX (0,0,1) model identified a positive and highly significant effect of PC1 with an 18-year latency ($\beta = 0.108$, $p < 0.001$). In prostate cancer, a model of identical structure showed an even larger and highly significant effect, with a 15-year latency ($\beta = 0.384$, $p < 0.001$). Both models passed all diagnostic checks, confirming statistical validity and robustness.

These findings offer robust quantitative evidence of long-latency relationships between animal product consumption and hormone-sensitive cancers. More broadly, the study highlights the relevance of econometric methodologies in cancer epidemiology and emphasizes their potential to deepen our understanding of how cumulative dietary exposures influence population health.

1. Introduction

In recent decades, the incidence of certain hormone-sensitive malignancies, such as breast and prostate cancer, has shown an exponential rise, in sharp contrast with the more linear increase observed for most other tumor types [1]. This divergence raises a crucial question: which factors make these diseases distinct from an epidemiological and biological perspective?

An interpretive key lies in their hormone-dependent nature. While the development and progression of these cancers are influenced by steroid hormones, it seems unlikely that the endogenous endocrine profile - remaining essentially stable across generations - alone could account for the exponential increase observed. It is therefore more plausible to hypothesize a substantial role for exogenous sources of hormonal stimulation: environmental or dietary substances capable of mimicking or amplifying physiological signals. Within this context, estrogens assume a central role. Their involvement in the etiopathogenesis of breast cancer is well documented [2][3][4], whereas in prostate cancer they have not traditionally been regarded as determining factors. However, recent evidence suggests that estrogens may act as early promoters of the neoplastic process, modulating the cellular and stromal environment during the initial phases of transformation, before yielding to the proliferative drive sustained by androgens [2][5][6]. Diet represents a primary route of exposure to bioactive molecules. Numerous epidemiological studies have suggested a link between meat and dairy consumption and a higher incidence of hormone-sensitive cancers [7][8][9]. However, most available investigations rely on time-limited observations, which are unable to capture the historical and cumulative dimension of such exposures. By contrast, dietary trends in Italy provide a unique opportunity: since the 1920s, ISTAT has collected detailed time series showing a sharp postwar increase in meat and dairy consumption, coinciding with the economic boom and the spread of industrialized diets. These historical discontinuities offer ideal conditions for exploring long-term causal relationships.

Within this framework, a major limitation of traditional studies on the diet–cancer relationship becomes evident: although they have highlighted important associations, they fail to adequately quantify the temporal dimension or the actual extent of latency - both crucial for correctly interpreting population-level phenomena. The analysis of aggregated time series could provide a privileged tool to bridge this gap, yet it is complicated by statistical challenges that may obscure or distort true relationships. The frequent non-stationarity of socio-economic and health data, together with biological delays that can extend over decades, creates conditions prone to spurious regressions - apparently significant associations between variables with similar trends but lacking any real causal link.

The present study aims to overcome these limitations by applying a rigorous econometric modeling framework to test and quantify the dynamic relationship between animal product consumption in Italy (1961–2020) and the incidence of two major hormone-sensitive cancers (1984–2020), breast and prostate. After verifying cointegration - demonstrating that dietary consumption trends and cancer incidence are not independent but linked by a long-term equilibrium - the analysis sought to estimate both the magnitude of the association and the number of years of latency with which such effects become statistically significant.

2. Results

2.1 Descriptive analysis

Annual time series for cancer incidence (*Breast, Prostate*) were considered from 1984 to 2020 (n = 37), while meat and dairy consumption were analyzed over a longer period, from 1961 to 2020 (n = 60), in order to account for potential lagged effects of dietary consumption on cancer incidence.

Log-transformed per capita meat consumption showed a mean of 3.94 and a standard deviation of 0.27, with values ranging from 3.16 to 4.21 (Table 1). Dairy consumption displayed a higher mean (5.42) and lower variability (standard deviation 0.19), indicating greater temporal stability compared to meat consumption (Table 1).

Table 1
Descriptive Statistics of the Time Series.

Variable	Temporal range	n	Mean	SD	Min	Max
Meat	1961–2020	60	3.94	0.27	3.16	4.21
Dairy	1961–2020	60	5.42	0.19	4.98	5.63
Consumption index PC1	1961–2020	60	0.00	1.38	-3.70	1.34
Breast	1984–2020	37	4.90	0.11	4.58	5.16
Prostate	1984–2020	37	4.71	0.31	3.91	5.07

Regarding the outcome variables, breast cancer incidence showed a mean of 4.90 and low variability (standard deviation 0.11), within a range of 4.58 to 5.16. The log-transformed incidence of prostate cancer was slightly lower (mean 4.71) but exhibited greater dispersion (standard deviation 0.31), with values ranging from 3.91 to 5.07, thus showing more pronounced fluctuations compared to breast cancer incidence.

The time series of meat and dairy consumption were found to be highly correlated at lag 0 (r = 0.9076), making the application of PCA necessary. As reported in Table 1S, PCA proved to be highly effective: the first principal component (PC1) explained more than 95% of the total variance of the original variables. The PC1 loadings, equal to 0.707 for both variables, were positive and of identical magnitude, confirming that PC1 represents meat and dairy consumption in a balanced way. This validates its interpretation as a composite index of animal product consumption, while avoiding informational redundancy in the models. By construction, the mean of PC1 is zero, while the standard deviation is 1.38, with values ranging from − 3.70 to 1.34 (Table 1).

Figure 1 shows the temporal trends of the time series under consideration and of the principal component PC1. To allow for direct visual comparison, all series were standardized (z-scores).

The graph reveals that PC1 exhibits a marked upward trend from 1960 until the late 1980s, followed by a plateau in the 1990s and a slight decline thereafter—an evolution similar to that observed in the original Dairy and Meat series.

Prostate cancer incidence increased until the early 2000s, after which it showed a slight decrease and subsequent stabilization. The growth phase and subsequent decline appear to be synchronized with variations in PC1, but with a temporal lag of more than a decade.

Breast cancer incidence, after a brief stabilization around 2010, showed instead a rapid and pronounced acceleration in the last decade, reaching an unusual peak around 2019–2020. Here too, the upward trend seems to follow that of PC1, with a substantial temporal lag.

In summary, the trajectory of the composite consumption index appears to visibly anticipate that of the cancer incidence curves. This visual evidence supports the hypothesis of a long-term relationship with structural lag, to be assessed through rigorous statistical tools (cointegration and ARIMAX) in order to avoid the risk of spurious regressions.

Step 1: Analysis of the Order of Integration and Cointegration Testing

The ADF tests indicated that all three series (Prostate, Breast, PC1) were non-stationary and integrated of order one (I(1)) (Table 2), thereby allowing the analysis to proceed to the subsequent steps.

Table 2
Augmented Dickey-Fuller tests in Levels and First Differences

	At-level time-series			First differences time-series		
Time Series	ADF Value I(0)	P value	Implication	ADF Value I(1)	P value	Implication
PC1	-1.5562	0.7542	Not stationary	-4.6175	0.0000	Stationary I(1)
Prostate	-0.0550	0.9900	Not stationary	-4.148	0.0000	Stationary I(1)
Breast	-1.9958	0.5749	Not stationary	-7.235	0.0000	Stationary I(1)

Before proceeding with ARIMAX-based dynamic modeling, the hypothesis of a cointegration relationship - namely, a long-term equilibrium between cancer incidence and the consumption index (PC1) - was tested using two approaches: Engle–Granger and the ARDL Bounds Test, to assess whether modeling in levels would be appropriate within the ARIMAX framework.. As reported in Table 3, the Engle-Granger test found no evidence of cointegration for either of the relationships examined ($p > 0.05$), likely due to the presence of short-term dynamics not captured by a static model, particularly in small samples. To overcome these limitations, the ARDL Bounds Test was applied (Table 3). This test, proposed by Pesaran et al. (2001), compares the calculated F-statistic with the critical values at the 5% significance level, with the null hypothesis (H_0) being the absence of cointegration. For the relationship Breast ~ PC1, the F-statistic (4.31) exceeded the upper 5% critical value (4.16), leading to rejection of H_0 and providing clear evidence of cointegration. The evidence was even stronger for the relationship Prostate ~ PC1, where the F-statistic (7.86) substantially exceeded the upper bound. Therefore, for both malignancies, the ARDL test confirmed the existence of a stable long-term relationship with the composite index PC1. This result provides robust validation for the use of ARIMAX models estimated in levels ($d = 0$).

Table 3
Cointegration Test Results for Breast and Prostate Cancer Incidence Relative to PC1 Consumption Index

Engle–Granger Cointegration Test					
Model Relationship	Test Statistic (τ)		p-value		Conclusion (α = 5%)
Breast ~ PC1	-1.8959		> 0.05		No cointegration detected
Prostate ~ PC1	-0.14417		> 0.05		No cointegration detected
ARDL Bounds Cointegration Test					
Model Relationship	Selected ARDL Model	F-Statistic	5% Critical Value Lower Bound (I(0))	5% Critical Value Upper Bound (I(1))	Conclusion
Breast ~ PC1	ARDL(1,0)	4.31	4.02	4.16	Cointegration confirmed
Prostate ~ PC1	ARDL(1,0)	8.26	4.02	4.16	Cointegration confirmed

Step 2: ARIMAX Framework

Step 2a: Identification of Structural Latency (lag L)

Before the definition of the optimal ARIMAX model, the temporal lag (L) of the delayed effect of PC1 on the outcome variables was first estimated. In line with the biological plausibility of carcinogenesis processes, lags within the [8–20]-year interval were considered, as this range was deemed reasonable for capturing potential long-term effects of dietary consumption on cancer incidence. To this end, preliminary ARIMAX models were estimated for each lag L within the defined interval. Figure 1S shows the variation of AICc as a function of the structural lag: for each model, the minimum of the curve identifies the optimal lag, representing the best compromise between goodness of fit and parsimony. For breast cancer, the lowest AICc value was observed at an 18-year lag, whereas for prostate cancer the minimum was found at 15 years. These values represent the optimal estimated latencies between changes in the consumption index (PC1) and cancer incidence trends, and were adopted in the final specification of the ARIMAX models.

Step 2b: Specification and Selection of the Best ARIMAX Model

Table 4 summarizes the results of the optimal model selection process for both malignancies. The choice was based on two hierarchical criteria:

Table 4
Comparative Analysis of ARIMAX Model Specifications for Breast and Prostate, based on Information Criteria and Residual Validation Tests

Model	ARIMAX(p,d,q) Model Specification	AICc	Q-Statistic (Ljung-Box)	p-value (Ljung-Box)	Shapiro–Wilk Statistic	p-value (Shapiro-Wilk)
Breast						
A	(2,0,0)	-78.08	5.48	> 0.05	0.9503	> 0.05
B	**(0,0,1)**	**-79.25**	**3.5295**	**> 0.05**	**0.9734**	**> 0.05**
C	(1,0,1)	-75.30	3.6988	> 0.05	0.9715	> 0.05
Prostate						
A	(1,0,0)	-48.17	11.05	> 0.05	0.9574	> 0.05
B	**(0,0,1)**	**-51.02**	**11.15**	**3.45**	**0.9677**	**> 0.05**
C	(1,0,1)	-48.65	7.18	2.08	0.9654	> 0.05

1. Statistical validity, verified through the Ljung–Box test (H_0: no residual autocorrelation) and the Shapiro-Wilk test (H_0: residuals are normally distributed);
2. Efficiency and parsimony, evaluated on the basis of corrected Akaike Information Criterion (AICc). minimization.

For breast cancer, three models were compared: Model A (ARIMAX(2,0,0)), Model B (ARIMAX(0,0,1)), both automatically selected by the algorithm, and Model C (ARIMAX(1,0,1)), specified manually. All three passed the diagnostic tests, proving statistically valid. The best model was identified as Model B, which had the lowest AICc value (–79.25).

Similarly, for prostate cancer, three models were tested: Model A (ARIMAX(1,0,0)), Model B (ARIMAX(0,0,1)), and Model C (ARIMAX(1,0,1)). All satisfied the statistical validity criteria. Here again, the optimal model was Model B, ARIMAX(0,0,1), as it showed the lowest AICc value (–51.02).

The selected ARIMAX models revealed a significant long-term relationship between PC1 - representing animal-based food consumption - and cancer incidence, as indicated by the significance of the predictor coefficients for both Breast and Prostate (Table 5).

Table 5
Parameter Estimates for Breast ARIMAX(0,0,1) and Prostate ARIMAX(0,0,1) Models (best models)

	Coefficient (β)	Standard Error	t-Statistic	p-value
Breast Model (Lag = 18) *ARIMA(0,0,1)*				
*PC1*_lagged (L = 18)	0.1084	0.0081	13.3318	< 0.001
MA(1) (Moving Average)	0.5493	0.1974	2.7833	< 0.01
Intercept	4.8986	0.0099	496.6983	< 0.001
Prostate Model (Lag = 15) *ARIMA(0,0,1)*				
*PC1*_lagged (L = 15)	0.3840	0.0203	18.8835	< 0.001
MA(1) (Moving Average)	0.7179	0.1366	2.8565	< 0.05
Intercept	4.6131	0.0193	238.7648	< 0.001

For breast cancer, the optimal model was an ARIMAX(0,0,1) with a structural lag of 18 years. The coefficient of the lagged PC1 variable was positive and highly significant (β = 0.1084, SE = 0.0081, p < 0.001), indicating that a 1% increase in PC1 corresponds, on average, to a 0.108% increase in breast cancer incidence. In addition to the long-term effect, the model also identified short-term dynamics, captured by a significant first-order moving average (MA(1)) term (β = 0.5493, p < 0.01).

For prostate cancer, a model of the same structure, ARIMAX(0,0,1), was selected with an optimal lag of 15 years. Here too, the coefficient of the lagged PC1 variable was positive and highly robust (β = 0.3840, SE = 0.0203, p < 0.001), corresponding to an elasticity of about 0.384. This implies that a 1% increase in PC1 is associated with an average 0.384% increase in prostate cancer incidence after 15 years. As with

breast cancer, the prostate cancer model also captured residual short-term dynamics through a significant MA(1) term ($\beta = 0.7179$, $p < 0.05$).

Overall, the results highlight a significant and temporally delayed influence of animal-based food consumption on the incidence of both cancers analyzed, with effects observable over a 15–18-year period. Importantly, for both models, the identification of a specification with no differencing ($d = 0$) was justified by the presence of cointegration, namely the existence of a long-term equilibrium between consumption and cancer incidence.

For breast cancer, the ARIMAX(0,0,1) model with an 18-year lag relative to the animal consumption index shows good agreement between the observed and estimated series, particularly up to the early 2000s: the forecasts closely follow the actual trend, capturing both the progressive increase in incidence and the slight declines. Short-term fluctuations are also well reproduced, confirming the appropriateness of the level specification (Fig. 3).

For prostate cancer, the ARIMAX(0,0,1) model with a 15-year lag provides an equally satisfactory fit, accurately reproducing the upward trend of the series. Discrepancies between observed and estimated values are minimal and show no systematic patterns, especially up to the early years of the new millennium (Fig. 3.)

Step 2c : Graphical Diagnostics of Model Residuals

Finally, the statistical validity of both selected ARIMAX models was also confirmed through graphical analysis of the residuals (Fig. 2S) for both breast and prostate cancer. The autocorrelation function (ACF) plots show that, in both cases, the residual correlations at different lags do not exceed the thresholds of statistical significance. In addition, the residuals fluctuate over time in a largely random manner around a zero mean, as illustrated by the time-series plot. Finally, the histograms of the residuals display unimodal and approximately symmetric frequency distributions, thereby visually supporting the assumption of normality of the errors, previously tested with the Shapiro–Wilk test (for both breast and prostate, $p > 0.05$).

3. Discussion

The results of this study show a significant long-latency relationship between animal product consumption and the incidence of breast and prostate cancer in Italy, with estimated temporal delays of 18 and 15 years, respectively. These findings, obtained through validated ARIMAX models supported by cointegration tests, indicate that changes in dietary consumption systematically precede variations in cancer incidence.

However, the mere observation of visual parallels between epidemiological trends and dietary consumption is not sufficient to demonstrate causality, as it may conceal the risk of spurious regressions arising from the non-stationarity of time series. For this reason, it was necessary to adopt a formal econometric approach capable of distinguishing long-term structural relationships from statistical noise and similar trends lacking biological meaning.

In the present study, the use of a rigorous framework combining ARDL Bounds Testing with ARIMAX modeling allowed us to validate the relationship between animal product consumption (summarized in the PC1 index) and the incidence of breast and prostate cancer. The methodological approach adopted is consistent with comparative studies on oncological data by the World Health Organization, which report the superiority of ARIMAX over traditional methods (joinpoint regression, AAPC) [19] [20].

For both malignancies, the results revealed a cointegration relationship with temporal lags of 18 years for breast cancer and 15 years for prostate cancer. These latencies are consistent with the minimum latency periods reported for cancers linked to environmental exposures, in line with multistage models of carcinogenesis [21][22]. This indicates that changes in dietary consumption systematically precede cancer incidence trends with a significant temporal delay, providing further evidence in support of a possible etiological link.

It is noteworthy that, despite similar latencies, the magnitude of the association differs markedly: more modest but statistically robust for breast cancer, and substantially higher for prostate cancer.

This difference may reflect a different biological sensitivity of the two malignancies to dietary hormone-mimetic stimuli, or interactions with disease-specific factors (such as the androgenic milieu for prostate cancer and estrogen-dependent proliferative mechanisms for breast cancer).

The biological interpretation of these results is consistent with the hypothesis that exogenous estrogens present in meat and dairy products may contribute to fostering a microenvironment favorable to carcinogenesis.

However, the model adopted does not allow for the isolation of the role of specific molecules or dietary components, nor does it distinguish the impact of consumption from that of other concomitant factors.

The main limitations stem from the use of aggregated national data. Ecological fallacy remains an intrinsic risk, since population-level associations cannot automatically be extended to individuals. Nonetheless, it has been observed that, for lifestyle-related cancers, aggregate and individual-level results tend to converge when exposure is relatively homogeneous across the population [23]. In addition, other potential confounding factors were not accounted for, such as changes in screening practices, therapeutic improvements, or demographic shifts, which may have influenced the observed trends.

In this context, particular attention should be given to the post-2000 dynamics, characterized by deviations from the expected trend. These variations should not be interpreted as a flaw of the model, but rather as the effect of unmodeled external factors that have altered the natural history of the cancers considered. For prostate cancer, the widespread use of PSA testing generated an artificial peak of diagnoses in the early 2000s, followed by a decline associated with its reduced use in clinical guidelines. For breast cancer, the stabilization observed in the early 2000s coincided with the sharp decrease in the use of hormone replacement therapy following the WHI trial (2002), while the subsequent acceleration in diagnoses can be attributed to the strengthening of screening programs and the introduction of technologies such as tomosynthesis.

Additional emerging risk factors must also be considered, including the increasing prevalence of obesity and changes in dietary patterns (greater consumption of ultra-processed foods, alcohol, and reduced fiber intake), which have likely reshaped the historical relationship between animal product consumption and cancer incidence.

Despite these complexities, the study presents elements of robustness and originality that strengthen its impact. The breadth of the time series (60 years of consumption and 37 years of incidence) represents a rare strength in cancer epidemiology. The rigorous handling of statistical challenges (non-stationarity, multicollinearity, autocorrelation) through PCA, ARDL, and ARIMAX reduces the risk of spurious results. The formal estimation of temporal lags (15–18 years) addresses a methodological gap in many traditional studies, while the comparative analysis of two hormone-sensitive cancers offers new insights into their differential biological vulnerability. Finally, the application of advanced econometric tools in the biomedical field represents an original methodological contribution capable of fostering interdisciplinary research.

Overall, the results not only reinforce the evidence of a long-term relationship between animal product consumption and hormone-sensitive cancers, but also demonstrate the ability of the models to capture structural changes related to public health interventions and lifestyle transformations, opening new perspectives for epidemiological interpretation.

In conclusion, this study provides robust statistical evidence supporting a long-term association, with latencies of 15–18 years, between animal product consumption and the incidence of breast and prostate cancer in Italy.

The analysis was based on long national time series and an advanced methodological framework combining PCA, cointegration tests, and ARIMAX models, which allowed for the rigorous quantification of both the magnitude and the temporal lag of the observed associations.

The deviations observed after 2000 should not be interpreted as a limitation of the model, but rather as evidence of its ability to capture structural changes attributable to the large-scale introduction of screening programs, the reduction in hormone replacement therapy use, and the emergence of new lifestyle- and obesity-related risk factors.

These findings suggest that chronic exposure to dietary sources of exogenous estrogens may represent a relevant determinant in the development of hormone-sensitive cancers. At the same time, the ability of the models to identify biologically plausible latencies and to detect epidemiological discontinuities reinforces their value as research tools capable of integrating consumption analysis with the interpretation of cancer trends.

4. Materials and Methods

4.1 Data Acquisition

All data used in this study were obtained from authoritative sources. Information on annual per capita food consumption (in kilograms) in Italy was retrieved from the FAO (Food and Agriculture Organization) database [10] [11]. Overall, the dataset covers the period 1961–2020, including both dairy products and meat.

Cancer incidence data for Italy were obtained from the ECIS (European Cancer Information System) platform [12], the official European Union resource providing epidemiological data on cancer for research purposes. The data collected represent annual age-standardized incidence rates (ASR per 100,000 population), calculated according to the age distribution of the European standard population. Overall, the dataset covers the period 1984–2020.

4.2 Dataset

To ensure that the data collected were consistent with the objectives of the study, specific statistical adjustments were performed. In particular, meat consumption was calculated by summing the FAO database categories “Bovine meat,” “Mutton & Goat meat,” and “Pigmeat.”

With regard to cancer data, a challenge arose from the way records are archived. In Italy, the oncological information available in the ECIS (European Cancer Information System) derives exclusively from regional or provincial registries, in the absence of a centralized national database. Consequently, in order to obtain a value representative of national incidence, the mean of the incidence rates reported by individual registries was calculated for each year.

Annual incidence data for prostate and breast cancers were denoted as Prostate and Breast, respectively. These two time series were treated as outcome variables in the analysis.

The primary explanatory variables were per capita meat consumption and per capita dairy consumption.

To avoid problems of heteroscedasticity, all time series were transformed into natural logarithms, thereby stabilizing variance and expressing the variables on a comparable scale.

Given the high correlation expected between Meat and Dairy consumption, including both as separate regressors in an econometric model could have introduced a serious problem of multicollinearity, making the coefficient estimates unstable and difficult to interpret. To address this issue, Principal Component Analysis (PCA) was employed - a dimensionality reduction technique that synthesizes the information contained in multiple correlated variables by using their mathematical transformations, while at the same time preserving the maximum possible variance, that is, the common information shared by the original variables [13].

In particular, a high Pearson correlation coefficient (r) between variables (e.g., $r > 0.7–0.8$) represents a strong indicator of the need to apply PCA. After confirming the high correlation between Meat and Dairy, the first principal component (PC1) was extracted and interpreted as a composite index of animal-based food consumption, encompassing both dairy and meat, and was used as an explanatory variable in the econometric models analyzed.

4.3 Study Design and Econometric Modeling Strategy

To investigate the dynamic relationship between the variables while mitigating the risk of spurious estimates - often encountered when analyzing non-stationary time series - a sequential and rigorous modeling strategy was adopted. The process, outlined in the flowchart shown in Fig. 2, was applied identically and independently to both malignancies (Prostate and Breast) and was structured in four stages to ensure the robustness and statistical validity of the final model:

Step 1: Analysis of the Order of Integration and Cointegration Testing

Step 2: ARIMAX Framework

Step 2a: Determination of Structural Lag

Step 2b: Specification and Selection of the ARIMAX Model

Step 2c: Graphical Diagnostics of the Residuals of the Selected Model

Step 1: Analysis of the Order of Integration and Cointegration Testing

As a preliminary step, the stationarity of the time series was assessed both in their original (level) form and after first differencing, in order to determine their order of integration, using the Augmented Dickey–Fuller (ADF) test[14]. The null hypothesis (H_0) of the test corresponds to the presence of a unit root (i.e., non-stationarity), whereas the alternative hypothesis indicates the absence of a unit root (i.e., stationarity). A time series that is stationary in levels is classified as I(0), while a series that becomes stationary only after first-order differencing is classified as I(1). Identifying I(1) series is a necessary condition for testing cointegration between cancer incidence and

dietary consumption. Cointegration, in fact, implies that although individual series are non-stationary, a long-term equilibrium exists that binds them together.

As a preliminary cointegration test, the Engle–Granger approach was adopted, which evaluates the null hypothesis of no cointegration by testing the stationarity of the residuals from a static OLS (Ordinary Least Squares) regression between the variables [15]. However, given the low power of this test in small samples, the results were regarded as merely exploratory and not conclusive, particularly in view of the limited length of the available time series.

To overcome this limitation, the ARDL Bounds Testing approach of Pesaran, Shin, and Smith (2001) was applied as the main test for cointegration [16]. This method, which is particularly robust in the presence of small samples, allows for the simultaneous estimation of both short- and long-term dynamics. The procedure involves estimating an Unrestricted Error Correction Model (UECM) and performing an F-test (Wald test) of the null hypothesis of no long-term relationship. The F-statistic is then compared against two sets of critical values (bounds): if the observed value exceeds the upper bound, the presence of cointegration is concluded.

Step 2: ARIMAX Framework

To model the relationship between cancer incidence and the index consumption PC1, the ARIMAX framework (Autoregressive Integrated Moving Average with eXogenous variables) was adopted. This approach was chosen because it simultaneously addresses three main challenges of the data: the inclusion of an external regressor (X), the handling of non-stationarity in the series, and the presence of short-term autocorrelation. The general form of the model is as follows:

$Y_t = \beta_0 + \beta_1 X_{t-L} + \eta_t$ (1)

where η_t is an error term following an ARIMA (p, d, q) process, designed to capture both non-stationarity (parameter d) and short-term dynamics (parameters p and q). The model was applied separately to the time series of Prostate and Breast incidence, using the animal consumption index PC1 as the exogenous egressor.

The identification of the optimal model specification, i.e., the orders p, d, and q, was carried out through three sequential steps:

- Step 2a: determination of the structural latency (lag L) with respect to the regressor PC1;
- Step 2b: selection of the most appropriate ARIMAX model based on AICc minimization and diagnostic validity;
- Step 2c: graphical diagnostics of the residuals of the selected model

All models were performed on the series in levels, setting d = 0, since the identification of the optimal model was possible without differenced series (d = 1), as prior cointegration testing (see Step 1) justified the use of undifferenced data.

Step 2a: Determination of Structural Lag (lag L)

In the first stage, the optimal temporal lag (L) of the delayed effect of PC1 on the outcome variables was identified within the 8–20-year interval, which was chosen as the analysis window in acknowledgment of the long timeframes required for carcinogenesis to develop. To this end, for each lag value a series of preliminary ARIMAX models was estimated using the auto.arima function from the forecast package. The optimal lag corresponded to the one associated with the model yielding the lowest Corrected Akaike Information Criterion (AICc) value. This criterion balances goodness of fit against model parsimony and is particularly suitable when dealing with small samples.

Step 2b: Specification and Selection of the ARIMAX Model

Once the optimal lag L was established, the structure of the ARIMAX model for each cancer type was determined, that is, the identification of the most appropriate p and q parameters to describe the temporal dynamics of the incidence series. To ensure the robustness of this choice, three different computational search strategies were employed, in order to select a set of three candidate ARIMAX models (Models A, B, and C) to be compared on the non-differenced series (d = 0).

Model A

identified using the auto.arima algorithm from the **forecast** package through a stepwise search of the best p and q parameters. The algorithm explores the parameter space efficiently, starting from an initial set of models and iteratively modifying one parameter at a time. Each modification is accepted only if it reduces the AICc, and the process stops when no further improvement is possible. Although computationally efficient, this approach does not necessarily guarantee the global optimum, with the risk of stopping at a local minimum.

To mitigate this risk, Model B was considered.

Model B

identified, again using auto.arima, through an exhaustive grid search over all possible combinations of p and q up to a predefined maximum order. Although more computationally demanding, this approach guarantees the identification of the model with the absolute minimum AICc, thereby eliminating the risk of converging to suboptimal solutions.

Model C

manually specified as an ARIMAX(1,0,1). This model was included for its parsimony and for its flexibility in representing different short-term dynamics.

The three models thus obtained were then compared and subjected to a thorough diagnostic assessment (described in Step 3) in order to evaluate their statistical consistency and select the final model.

Finally, the choice of the final model was made according to two hierarchical criteria:

i. Statistical validity, and
ii. Efficiency and parsimony.

With respect to statistical validity, the residuals (ϵ_t) of each model were required to satisfy the properties of a white noise process (zero mean, constant variance, and no autocorrelation). This condition was verified using the Ljung–Box test, whose null hypothesis assumes the absence of serial autocorrelation [17]. As an additional diagnostic check, residual normality was also assessed using the Shapiro–Wilk test [18]. Only models for which both tests returned a p-value > 0.05 were considered valid.

The second criterion, related to efficiency and parsimony, required the selection—among statistically valid models—of the one with the lowest AICc value.

Step 2c: Graphical Diagnostics of the Residuals of the Selected Model

The adequacy of the ARIMAX model identified as the best was further assessed through a graphical analysis of the residuals. Specifically, the following aspects were examined:

i. the temporal pattern of the residuals, which was expected to be free of systematic structures and to display random behavior;
ii. the autocorrelation function (ACF), in order to confirm the absence of residual serial dependence;
iii. the distribution of the residuals through a histogram, to evaluate their compatibility with normal dric.

All statistical analyses were performed in the **R** environment (version 4.x.x), using the *forecast t-series* packages.

tep 2c: Graphical Diagnostics of the Residuals of the Selected Model

The adequacy of the best-fitting ARIMAX model was further assessed through graphical diagnostics of the residuals. Specifically, the following aspects were evaluated:

i. the time-series plot of the residuals, which was expected to exhibit random fluctuations with no systematic patterns;
ii. the autocorrelation plot (ACF), used to verify the absence of residual serial dependence;
iii. the histogram of residuals, used to assess their adherence to a normal distribution.

All statistical analyses were conducted in the R environment [24], using the forecast and tseries packages.

Declarations

6. Acknowledgements

The authors thank Lucia Dentale for her support with the translation.

7. Funding Declaration

This research received no external funding.

8. Author contributions

Conceptualization: all authors; Writing – original draft: A.S., M.T., N.C., G.E.R., and A.T.; Writing – review and editing: A.S., M.T., E.P.A., N.C., G.E.R., R.I., P.I., and A.T.; Visualization: A.S., E.P.A, and P.I.; Supervision: A.S., G.E.R., and A.T. All authors have read and approved the final version of the manuscript.

9. Data availability statement

The datasets analyzed during the current study are publicly available. Historical food balance sheets were obtained from the Food and Agriculture Organization of the United Nations (FAOSTAT) and updated food balance sheets were accessed from FAOSTAT. Cancer incidence and mortality data were retrieved from the European Cancer Information System (ECIS) of the European Commission. All datasets are openly accessible through the respective repositories at the following links:

a. FAOSTAT historical (https://www.fao.org/faostat/en/#data/FBSH)

b. FAOSTAT food balance sheets (https://www.fao.org/faostat/en/#data/FBS),

c. ECIS data explorer (https://ecis.jrc.ec.europa.eu/data-explorer#/historical/incidence-mortality-by-cancer?ageFrom=0&ageTo=85%2B&indicator=IN&sex=0&yearFrom=1976&yearTo=2015&cancerEntity=-1&statistic=ASR_EU_NEW&registry=127)

Conflicts of Interest statement

The authors declare no conflict of interest.

References

1. Online summary of trends in U.S. cancer control measures. *National Cancer Institute Cancer Trends Progress Report*https://progressreport.cancer.gov/diagnosis/incidence
2. Cavalieri, E., Rogan, E. The 3,4-quinones of estrone and estradiol are the initiators of cancer whereas resveratrol and N-acetylcysteine are the preventers. *Int. J. Mol. Sci*. 2021 Jul 30;22(15):8238. doi: 10.3390/ijms22158238. PMID: 34361004; PMCID: PMC8347442.
3. Yager, J. D., Davidson, N. E., Estrogen carcinogenesis in breast cancer. *N. Engl. J. Med.* 2006 Jan 19;354(3):270-82. doi: 10.1056/NEJMra050776. PMID: 16421368.
4. Liehr, J. G., Is estradiol a genotoxic mutagenic carcinogen? *Endocr. Rev.* 2000 Feb;21(1):40-54. doi: 10.1210/edrv.21.1.0386. PMID: 10696569.
5. Ozten, N. et al. Role of estrogen in androgen-induced prostate carcinogenesis in NBL rats. *Horm. Cancer* 2019 Jun;10(2-3):77-88. doi: 10.1007/s12672-019-00360-7. Epub 2019 Mar 16. PMID: 30877616; PMCID: PMC6545235.
6. Rahman, H. P., Hofland, J., Foster, P.A. In touch with your feminine side: how oestrogen metabolism impacts prostate cancer. *Endocr. Relat. Cancer* 2016 Jun;23(6):R249-66. doi: 10.1530/ERC-16-0118. Epub 2016 May 18. PMID: 27194038.
7. Zhang, J., Kesteloot, H. Milk consumption in relation to incidence of prostate, breast, colon, and rectal cancers: is there an independent effect? *Nutr. Cancer* 2005;53(1):65-72. doi: 10.1207/s15327914nc5301_8. PMID: 16351508.
8. Besson, H., Paccaud, F., Marques-Vidal, P.. Ecologic correlations of selected food groups with disease incidence and mortality in Switzerland. *J. Epidemiol.* 2013;23(6):466-73. doi: 10.2188/jea.je20130029. Epub 2013 Oct 19. PMID: 24140818; PMCID: PMC3834285.
9. Grasgruber, P., Hrazdira, E., Sebera, M., Kalina, T. Cancer incidence in europe: an ecological analysis of nutritional and other environmental factors. *Front. Oncol.* 2018 Jun 13;8:151. doi: 10.3389/fonc.2018.00151. PMID: 29951370; PMCID: PMC6008386.
10. FAOSTAT: food balance sheets historical. *Food and Agriculture Organization of the United Nations*https://www.fao.org/faostat/en/#data/FBSH
11. FAOSTAT: food balance sheets. *Food and Agriculture Organization of the United Nations*https://www.fao.org/faostat/en/#data/FBS
12. European Cancer Information System (ECIS). *European Commission*https://ecis.jrc.ec.europa.eu/data-explorer#/historical/incidence-mortality-by-cancer?ageFrom=0&ageTo=85%2B&indicator=IN&sex=0&yearFrom=1976&yearTo=2015&cancerEntity=-1&statistic=ASR_EU_NEW&registry=127
13. Greenacre, M. et al. Principal component analysis. *Nat. Rev. Methods Primers* 2, 100 (2022). https://doi.org/10.1038/s43586-022-00184-w
14. Paraproditis, E., Politis, D. N. The asymptotic size and power of the augmented Dickey–Fuller test for a unit root. *Econometric Reviews* 37, 955–973 (2018).

15. Engle, R. F., Granger, C. W. J. (1987). Co-Integration and error correction: representation, estimation, and testing. *Econometrica* 55(2), 251–276. https://doi.org/10.2307/1913236
16. Pesaran, M. H., Shin, Y., Smith, R. J. (2001) Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics* 16, 289-326. http://dx.doi.org/10.1002/jae.61
17. Ljung, G. M., Box, G. E. P. On a measure of lack of fit in time series models, *Biometrika*, Volume 65, Issue 2, August 1978, Pages 297–303, https://doi.org/10.1093/biomet/65.2.297
18. Shapiro, S. S., Wilk, M. B. An analysis of variance test for normality (complete samples), *Biometrika*, Volume 52, Issue 3-4, December 1965, Pages 591–611, https://doi.org/10.1093/biomet/52.3-4.591
19. Li, J., Chan, N. B., Xue, J., & Tsoi, K. K. (2022). Time series models show comparable projection performance with joinpoint regression: A comparison using historical cancer data from World Health Organization. *Frontiers in Public Health10*, 1003162.
20. Trächsel, B., Rousson, V., Bulliard, J. L., & Locatelli, I. (2023). Comparison of statistical models to predict age-standardized cancer incidence in Switzerland. *Biometrical journal. Biometrische Zeitschrift*, *65*(7), e2200046. https://doi.org/10.1002/bimj.202200046
21. Nadler, D. L., & Zurbenko, I. G. (2014). Estimating cancer latency times using a Weibull model. *Advances in Epidemiology* 2014(1), 746769.
22. Little, M. P., Eidemüller, M., Kaiser, J. C., & Apostoaei, A. I. (2024). Minimum latency effects for cancer associated with exposures to radiation or other carcinogens. *British journal of cancer* 130(5), 819–829. https://doi.org/10.1038/s41416-023-02544-z
23. Lokar, K., Zagar, T., & Zadnik, V. (2019). Estimation of the ecological fallacy in the geographical analysis of the association of socio-economic deprivation and cancer incidence. *International Journal of Environmental Research and Public Health* 16(3), 296.Inizio modulo
24. Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA.

Figures



Figure 1

Standardized trends (z-scores) of meat consumption, dairy consumption, animal product consumption index (PC1), breast cancer incidence, and prostate cancer incidence (1961–2020).

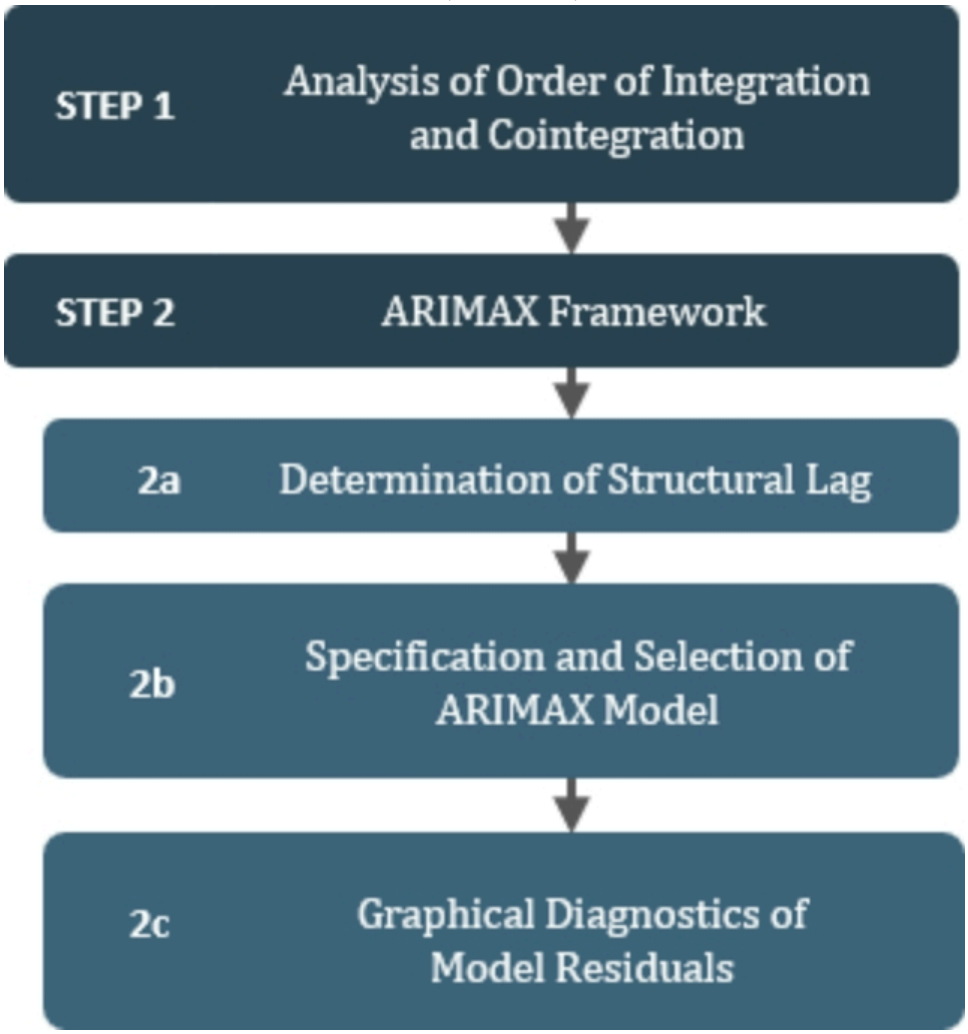


Figure 2

Flowchart of Time-Series Analysis

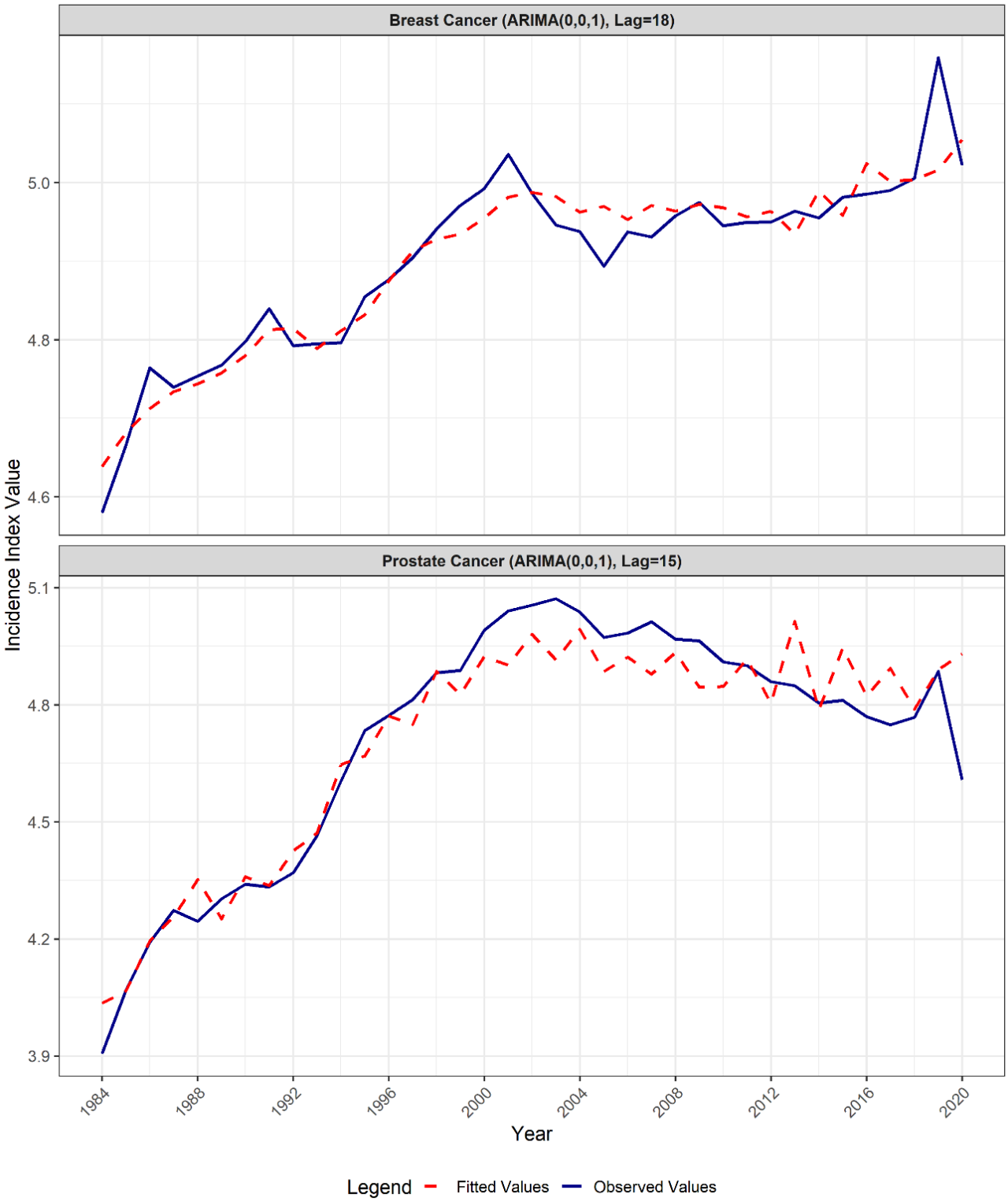


Figure 3

Fitted Values for the Final Selected Models: Breast Cancer ARIMAX(0,0,1) model with lag=18; Prostate Cancer ARIMAX(0,0,1) model with lag=15. vs observed values

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx