

Supplementary Table 1 Data Sources and Extraction Details for *H. pylori* Infection Rates

Table 1 Source Literature

Reference	First Author	Year	Journal	Study Type	Sample Size	Coverage	Primary Use
[41]	Li M	2020	Helicobacter	Systematic review and meta-analysis	670,572 (98 studies)	26 provinces, 1983–2018	Urban-rural stratification framework, Meta-analysis framework
[42]	Wang Z	2024	Helicobacter	Surveillance system study	1,267,988 (384 surveys)	31 provinces, 1983–2019	Temporal trend data, Detection method correction

Table 2 Literature Search Strategy (from Li et al [41] and Wang et al [42])

Item	Li et al [41]	Wang et al [42]
Databases	PubMed, Embase, CNKI, Wanfang	Validated Li et al [41] search strategy + Supplementary search for Hong Kong, Macao, and Taiwan

Search period	January 1, 1983 - June 4, 2020	1990–2019 (Based on Li et al [41] extension)
Search terms	" <i>H. pylori</i> " / " <i>H. pylori</i> " / " <i>HP</i> " / " <i>Campylobacter pylori</i> " + "prevalence" + "China"	Same as Li et al [41], Supplemented with Hong Kong, Macao, and Taiwan terms
Language restriction	English, Chinese	English, Chinese
Supplementary search	Manual search of reference lists	Supplementary search for Hong Kong, Macao, and Taiwan reports
Quality assessment	Hoy et al. risk of bias tool (10 items, ≥ 6 = low bias)	Loney criteria

Table 3 Inclusion and Exclusion Criteria (from Li et al [41])

Inclusion criteria:

1. (1) Original studies reporting *H. pylori* infection rates in adults (≥ 18 years) from mainland China;
2. (2) Representative of the general population;
3. (3) Clearly reported study location, time period, and detection method.

Exclusion criteria:

4. (a) Non-original peer-reviewed studies (reviews, letters, conference abstracts, etc.);
5. (b) Studies exclusively focusing on specific populations (elderly, migrant populations, patients with digestive diseases, etc.);
6. (c) Missing key population characteristics (age, sex, location, time period);
7. (d) Subset of another study;
8. (e) Unspecified detection method or non-gastric infection status tests (e.g., saliva test);

9. (f) Only reported data for children;

10. (g) Unknown sample size.

Table 4 H. pylori Infection Rates Used in This Study (Calibrated Based on Urban-Rural Rates Reported in Literature for Each Time Period)

Time Window	National Rate (%)	Urban Rate (%)	Rural Rate (%)	Urban-Rural Ratio (Rural/Urban)	Primary Data Source
1989–1994	63.8	50.6	77.0	1.52	Li et al [41] 1983–1994 period + Literature calibration for urban- rural rates
1995–1999	59.3	47.8	70.8	1.48	Li et al [41] 1995–2005 period + Literature calibration for urban- rural rates
2000–2004	54.8	46.3	63.3	1.37	Wang et al [42] 2000– 2009 data + Literature calibration

					for urban-rural rates
2005–2009	50.3	45.7	54.9	1.20	Wang et al [42] 2000–2009/ 2010–2014 data + Literature calibration
2010–2014	43.4	38.6	48.1	1.25	Wang et al [42] 2010–2014 data + Literature calibration for urban-rural rates
2015–2018	38.7	30.8	45.4	1.47	Wang et al [42] 2015–2019 data + Literature calibration for urban-rural rates

Note: Urban and rural infection rates were calibrated based on actual survey data reported in the literature for each time period, rather than using a fixed ratio. The urban-rural ratio changes over time, reflecting the dynamic changes in urban-rural disparities across different historical periods.

Table 5 Trends in Urban-Rural Ratio

Time Window	Urban-Rural Ratio (Rural/Urban)	Trend
1989–1994	1.52	Period of largest urban-rural disparity
1995–1999	1.48	Disparity slightly narrowed
2000–2004	1.37	Disparity continued to narrow
2005–2009	1.20	Period of smallest disparity (improved rural sanitation)
2010–2014	1.25	Disparity slightly rebounded
2015–2018	1.47	Disparity expanded again (may reflect detection method differences)

Note: The lowest urban-rural ratio in 2005–2009 (1.20) may be associated with the "Home Appliances Going to the Countryside" policy, improved rural drinking water, and other public health measures. The rebound in 2015–2018 may be related to the continued use of serological tests in rural areas (which overestimate infection rates).

Table 6 Distribution and Calibration of Detection Methods (from Wang et al [42])

Time Window	Primary Detection Method	Urban Detection Method Characteristics	Rural Detection Method Characteristics	Calibration Note

1989–1994	Serology (IgG) predominant	Predominantly serology	Predominantly serology	Serology may overestimate infection rate (antibodies persist)
1995–1999	Serology predominant	Predominantly serology	Predominantly serology	Same as above
2000–2004	Serology → UBT transition	UBT gradually popularized	Still predominantly serology	Urban-rural detection method differences began to emerge
2005–2009	UBT gradually popularized	UBT predominant	Serology/UBT coexist	Urban-rural disparity narrowed
2010–2014	UBT predominant	UBT predominant	UBT gradually popularized	Detection methods became more consistent
2015–2018	UBT standardized	UBT standardized	UBT increased but serology still present	Higher proportion of serology in rural areas

Note: Wang et al [42] used the formula $\text{prev} = (\text{rate} + \text{spe} - 1) / (\text{sen} + \text{spe} - 1) \times 100$ to correct for different detection methods. Li et al [41] did not perform such correction, and their reported urban-rural infection rates may be affected by detection method differences.

Table 7 Original Urban-Rural Stratification Data from Li et al [41] Meta-

Analysis

Category	No. Studies	Sample Size	Pooled Infection Rate (% , 95% CI)
Urban	62	389,913	44.6 (41.7, 47.6)
Rural	24	257,782	59.2 (54.7, 63.5)
Unspecified	12	22,877	56.7 (48.2, 64.7)
Overall urban-rural ratio	—	—	1.33 (OR=1.80, 95% CI: 1.43–2.27)

Note: The overall urban-rural ratio from Li et al [41] (1.33) is lower than the actual ratios in each time period of this study (1.20–1.52), indicating that using a fixed ratio would underestimate early urban-rural disparities and overestimate recent disparities.

Table 8 Key Sources of Heterogeneity (from Li et al [41] Moderator Analysis)

Variable	τ^2	R ² (%)	OR (95% CI)	P-value
Study period (ref: 1983–1994)	0.2508	18.8		
1995–2005			0.77 (0.49, 1.20)	0.24
2006–2018			0.50 (0.36, 0.69)	<0.0001
Urban/rural status (ref: Urban)	0.2370	24.0		
Rural			1.80 (1.43, 2.27)	<0.0001
Detection method (ref:	0.2539	17.8		

UBT)				
IgG serology			1.73 (1.33, 2.26)	<0.0001
Study design (ref: Population-based)	0.2149	30.4		
Hospital-based			0.53 (0.44, 0.65)	<0.0001

Table 9 Comparison of Temporal Trends

Data Source	Period	Urban Rate Change	Rural Rate Change	National Rate Change
Li et al [41]	1983–1994 → 2006–2018	—	—	63.8% → 46.7% (↓27%)
Wang et al [42]	1983–1999 → 2015–2019	—	—	63.3% → 38.7% (↓39%)
This study	1989–1994 → 2015–2018	50.6% → 30.8% (↓39%)	77.0% → 45.4% (↓41%)	63.8% → 38.7% (↓39%)

Table 10 Limitations of Data in This Study

Limitation	Description
Non-continuous surveillance data	Both sources are literature compilations, not national continuous surveillance
Time window splitting	Wang et al [42] had 4 windows that needed to be split into 6 windows for this study; some periods required interpolation
Uneven urban-rural literature	Fewer rural surveys than urban (Li et

distribution	al [41]: 62 urban vs. 24 rural studies)
Detection method differences	Rural areas predominantly used serology in early periods, which may overestimate infection rates; Wang et al [42] corrected for this, but historical data remain uncertain
Limited calibration basis	Urban-rural calibration for some periods relied on limited literature reports with potentially limited sample representativeness

Supplementary material 2 Full Reproducible Methods Document

2.1 Study dataset and analytical time window

2.1.1 Outcome: Gastric cancer incidence data

Analysis time window for correlation analysis: 1989-2018 (observed registry data only). 2022: BAPC-model projected estimates, which were excluded from all bivariate correlation calculations and only used for trend description in the main text.

Data source: Gastric cancer registry database from the National Cancer Center (NCC) of China.

Indicator: ASIRC – Age-standardized incidence rate by Chinese standard population (per 100 000 person-years).

Stratification: Stratified by urban/rural residence and sex (male / female).

Unit: Cases per 100 000 person-years.

2.1.2 Explanatory variables dataset

Data sources: *China Statistical Yearbook*, environmental-protection statistical yearbooks, peer-reviewed published literature, and national health statistical yearbooks. All variables are annual time-series variables spanning 1989-2018. The categories of included covariates are listed below:

- Economic variables: Annual per-capita disposable income; annual ownership of refrigerators per 100 households. - Dietary variables: Annual per-capita fresh-fruit consumption; annual per-capita fresh-vegetable consumption; daily per-capita intake of high-salt preserved foods; daily per-capita processed-meat consumption; average daily per-capita alcohol consumption; smoking proportion. - Environmental-pollution variables - Industrial wastewater: Annual chemical oxygen demand (COD) emissions; annual ammonia-nitrogen emissions - Municipal domestic sewage: Annual chemical oxygen demand (COD) emissions; annual ammonia-nitrogen emissions - Demographic-migration variables: Proportion of urban population; proportion of rural population. - *H. pylori* prevalence: Extracted from segmented published literature reviews (no nationwide continuous surveillance dataset available).

Unit notes: Monetary variables were adjusted to constant prices to enable comparability across years. Pollutant-emission unit: tons per year.

Consumption-related variables retain the original units documented in official yearbooks: kg/capita/year or g/capita/day.

2.2 Data pre-processing steps

2.2.1 Unit harmonization & data transformation

All calendar years were aligned to the period 1989-2018. Monetary indicators were inflation-adjusted using the national consumer price index (CPI) to eliminate price-related bias across different years. For dietary consumption variables, unit conversions between annual and daily values were performed (annual consumption divided by 365 to calculate daily per-capita intake). Pollutant-emission values retained their original statistical unit (tons per year); no additional population-based normalization was applied in primary bivariate analyses.

2.2.2 Missing-data handling

Each bivariate correlation was computed using only pairwise complete time-points, without global imputation. The effective sample size n for each correlation equals the number of non-missing paired time-points for gastric-cancer incidence and the corresponding explanatory variable. Missing values were neither interpolated nor filled. Exact sample size n and study-year range for each variable are reported in the footnote of Table 1 in the main manuscript.

2.2.3 Time-alignment procedure

Outcome and predictor variables were matched strictly by calendar year. No time-point shifting was implemented to account for latency effects in primary bivariate analyses.

Note: Latency effects are acknowledged as an important limitation in the main text; formal latency-lag time-series modelling was not performed in this study.

2.3 H. pylori prevalence: See Supplementary Table S1 for full details

2.4 Joinpoint regression analysis parameters

Joinpoint software version: Joinpoint Regression Program v4.9.1.0 (National Cancer Institute, United States).

Dependent variable: Age-standardized gastric-cancer incidence rate (ASIRC, per 100 000 person-years).

Stratification: Rural male, urban male, rural female, urban female.

Model settings: - Minimum number of observations between joinpoints: 4 - Maximum permitted number of joinpoints: 3 - Error-variance assumption: Poisson distribution - Output metrics: Annual Percent Change (APC), Average Annual Percent Change (AAPC), together with 95% confidence intervals.

Input time range: Observed registry data from 1989-2018. BAPC-projected estimates for 2022 were used only for trend visualization and were not incorporated into Joinpoint model fitting.

Significance criterion for joinpoints: $P < 0.05$.

Output: APC, AAPC and their corresponding 95% confidence intervals are presented in the main-text results section.

2.5 Bivariate ecological correlation analysis workflow

All correlation analyses were performed exclusively on observed 1989-2018 incidence data; BAPC-projected 2022 estimates were excluded.

Normality testing: The Shapiro-Wilk test was applied to each paired time-series dataset.

Correlation-method decision rules: - Pearson correlation coefficient was adopted when the bivariate-normality assumption was satisfied. - Spearman rank-order correlation coefficient was adopted for non-normal distributions.

Statistical outputs: Correlation coefficient r , 95% confidence interval, two-tailed P -value.

Interpretation of P -values: Reported two-tailed $P < 0.05$ serves only descriptive purposes; no formal correction for multiple-testing / multiplicity was performed.

Descriptive ranking: After excluding predefined confounder candidates, the coefficient of determination R^2 was calculated purely for descriptive ranking.

Critical note: R^2 values reflect unadjusted bivariate temporal associations only and cannot quantify causal strength or the magnitude of independent effects.

Important acknowledged limitations of bivariate correlation: Simple bivariate correlation cannot account for secular trends, temporal autocorrelation, latency periods, confounding effects, or multiplicity bias. All outputs

represent exploratory descriptive temporal associations; causal inference cannot be drawn. Time-series variables with co-varying trends are susceptible to spurious associations.

2.6 BAPC model brief note

The BAPC model was used to project gastric-cancer incidence estimates for 2020-2022. Projected estimates were used solely for trend description and were never entered into correlation analyses or Joinpoint regression. Model parameters follow those described in the original BAPC reference publication. Full BAPC configuration settings are archived together with analytical files.

2.7 Statistical code and output availability

Complete SPSS statistical outputs and the corresponding input-data Excel (xlsx) file are provided.