Original Paper

Viability of Public Statistics–Based Synthetic Personas in Small-Area Public Health Surveillance: Sociodemographic Benchmark Consistency Study Using Korea Community Health Survey and Korean Statistical Information Service Data

Miryoung Kim, PhD

College of Pharmacy, Sunchon National University, Suncheon, Jeollanam-do, Republic of Korea

Corresponding Author:
Miryoung Kim, PhD
College of Pharmacy
Sunchon National University
255 Jungang-ro
Suncheon, Jeollanam-do, 57922
Republic of Korea
Phone: 82 61 750 3763
Email: miryoung.kim@scnu.ac.kr

Abstract

Background: Synthetic population and persona datasets can provide artificial individual-level records without exposing personal data. However, for surveillance use, they must preserve the joint profiles that define vulnerable groups, not only marginal distributions.

Objective: This study evaluated whether a public statistics–based Korean synthetic persona dataset is consistent with Korea Community Health Survey (CHS) and Korean Statistical Information Service (KOSIS) benchmarks for auxiliary small-area public health surveillance.

Methods: We compared 1,000,000 synthetic persona rows with 1,156,122 adult respondents from the 2021-2025 CHS and selected KOSIS municipal indicators. Core variables were age group, sex, marital status, living-alone status, and education; province or municipal area replaced marital status in the regional profiles. Economic activity and occupation were treated as sensitivity domains. We compared marginal and joint distributions using percentage-point differences and total variation distance (TVD) and quantified CHS sampling uncertainty using 500-replicate stratified cluster percentile bootstrap intervals and compared high-dimensional TVD with CHS split-half references. Sparse-cell, KOSIS correlation, and regional composition-standardized analyses assessed robustness and potential downstream use.

Results: Persona rows were close to CHS for sex (TVD<0.001, 95% bootstrap percentile interval 0.000-0.001), province (TVD 0.006, bootstrap percentile interval 0.005-0.006), living-alone status (TVD 0.008, bootstrap percentile interval 0.007-0.009), and age group (TVD 0.013, bootstrap percentile interval 0.012-0.014). Differences were larger for marital status (TVD 0.039, bootstrap percentile interval 0.038-0.040) and education (TVD 0.059, bootstrap percentile interval 0.058-0.061). The national core profile had a TVD of 0.106 (bootstrap percentile interval 0.105-0.107), compared with a CHS split-half median of 0.009 (2.5th-97.5th percentile 0.008-0.011). The municipal profile had an observed TVD of 0.117, a bootstrap mean of 0.122 (95% bootstrap percentile interval 0.120-0.123), and a split-half median of 0.082 (2.5th-97.5th percentile 0.080-0.084). After restriction to CHS cells with n≥100, the municipal TVD decreased to 0.091. At the municipal level, CHS- and persona-composition standardized unmet health care rates differed by a mean of 0.109 percentage points (Pearson r=0.567), but standardization with national profile-specific probabilities did not reproduce observed local CHS rates (Pearson r=0.004 for persona composition). The conditional KOSIS living-alone comparison was modest (Pearson r=0.547).

Conclusions: The synthetic persona data showed close agreement for several basic demographic margins but less consistent agreement for high-dimensional sociodemographic profiles. They may support composition-based scenario development and dashboard prototyping but should not be used as stand-alone sources for local surveillance or health-outcome inference without additional calibration and validation.

XSL•FO
RenderX

(JMIR Public Health Surveill 2026;12:e106306) doi: 10.2196/106306

KEYWORDS

digital public health; public health informatics; public health surveillance; synthetic population; persona; small-area analysis; benchmark validation; Korea community health survey; KOSIS; Korean Statistical Information Service; health equity

Introduction

Effective small-area public health surveillance requires more than estimating average health indicators by region [1,2]. Decision-makers also need to know which groups are vulnerable, where they live, and how demographic and socioeconomic characteristics cluster within local populations [1,3]. These profiles are difficult to observe reliably at small-area levels because survey data become sparse after stratification, whereas public statistics–based systems commonly provide marginal distributions or selected cross-tabulations rather than flexible individual-level structures [3,4].

Synthetic population and persona datasets are therefore attractive for public health informatics because they can provide artificial individual-level records for scenario analysis, vulnerability profiling, dashboard development, and data-system testing without exposing real personal records [5-8]. Their value for surveillance, however, depends on whether they preserve the joint profiles that define vulnerability, not simply whether they look plausible at the margin [9-13]. Prior work on synthetic patients, synthetic microdata, and population synthesis has emphasized that resemblance, utility, privacy, and intended-use validity are distinct evaluation dimensions, and that fitting marginal totals may still distort multivariable dependence structures [6-14].

This issue has become timely in Korea because Nemotron-Personas-Korea, a public statistics–based Korean synthetic persona dataset, is now available for reuse in digital population health workflows. The dataset was released by NVIDIA under a Creative Commons Attribution (CC-BY) 4.0 license and includes 1 million rows covering 17 provinces and 252 districts, with demographic, household, educational, occupational, and regional attributes [15]. The dataset documentation describes seed sources including Korean Statistical Information Service (KOSIS) census and public statistics data; Supreme Court data on name, birth year, and sex; National Health Insurance Service health checkup information; Korea Rural Economic Institute food consumption behavior survey materials; and NAVER Cloud seed data and domain expertise [15]. However, public availability and broad demographic coverage do not establish fitness for surveillance use.

The Korean setting provides a useful benchmark structure. The Korea Community Health Survey (CHS) produces observed small-area health indicators for local health planning and monitoring at the city, county, and district levels [16,17]. Synthetic personas should therefore not be presented as substitutes for CHS municipal-level estimates. A more appropriate question is whether they can serve as an auxiliary analytic layer for scenario analysis, vulnerability profiling, dashboard prototyping, and calibration. CHS individual records provide a complex-sample survey benchmark for observed combinations of individual characteristics [16,17], whereas KOSIS provides seed-source–informed public statistics benchmarks for selected regional characteristics [18].

This study evaluates whether a public statistics–based Korean synthetic persona dataset is consistent with CHS and KOSIS benchmarks in ways that matter for auxiliary small-area public health surveillance. We use a layered validation approach that moves from basic margins to joint distributions, high-dimensional profiles, policy-relevant vulnerable profiles, sparse-cell sensitivity, CHS temporal stability, KOSIS regional ranking consistency, split-half noise references, independence baselines, and downstream composition-standardized sensitivity. The purpose is to distinguish where synthetic personas can responsibly support descriptive planning or scenario-oriented work and where additional calibration is needed before subgroup or small-area inference [9,12,13].

Methods

Study Design

We performed a benchmark consistency study of a synthetic persona dataset intended for auxiliary public health informatics and surveillance planning [12,13]. The primary aim was not to predict CHS municipal-level unmet health care rates, but to assess whether synthetic persona rows reproduced sociodemographic distributions observed in CHS [16,17]. We compared persona distributions with CHS complex-survey benchmarks and selected KOSIS aggregate indicators that also informed persona generation [18]. Table 1 summarizes the benchmark comparison layers, data sources, analytic metrics, and intended interpretation.

XSL•FO
RenderX

Table 1. Validation layers, benchmark sources, metrics, and interpretation[a].

Validation layer	Benchmark source	Main metrics	Surveillance interpretation	Main limitation
Marginal distributions	CHS[b]	TVD[c], JS[d] divergence, and percentage-point difference	Basic sociodemographic structure check	Does not assess whether characteristics co-occur correctly
Joint and high-dimensional profiles	CHS	TVD, percentile bootstrap interval, CHS split-half TVD, and independence baseline TVD	Evaluates whether profile structure exceeds margin-only agreement	Sparse cells and municipal instability limit direct subgroup inference
Policy-relevant vulnerable profiles	CHS	Weighted percentage, difference in percentage points, and CHS 95% CI	Interpretable profile-level consistency for planning scenarios	Profile consistency does not validate persona-based health outcomes
KOSIS[e] public-statistics benchmarks	KOSIS	Pearson and Spearman correlations and exact-match sensitivity	Seed-source–informed aggregate consistency check	Not fully independent; denominators and source years differ
Sensitivity domains	CHS and KOSIS	Economic activity, occupation, KOSIS exact-match, and sparse-cell diagnostics	Assesses robustness to harmonization and sparse-cell decisions	Exploratory; not used as core evidence

[a]Each layer evaluates a different aspect of benchmark consistency rather than replacing official CHS or KOSIS estimates.

[b]CHS: Korea Community Health Survey.

[c]TVD: total variation distance.

[d]JS Divergence: Jensen-Shannon divergence.

[e]KOSIS: Korean Statistical Information Service.

Data Sources and Benchmark Roles

The synthetic persona dataset was treated as the validation target. We used NVIDIA Nemotron-Personas-Korea, an open-source Korean-language synthetic persona dataset released under a CC-BY 4.0 license [15]. Version 1.0 was released on April 20, 2026. The dataset card published by NVIDIA on Hugging Face describes 1 million rows, with structured fields including sex, age, marital status, family type, education level, occupation, district, province, and country, as well as 7 narrative persona text fields [15]. The analytic unit in this study was the row-level structured demographic record, not the narrative text fields. The dataset card identifies the following sources: KOSIS census tables covering sex, province, industry, occupation, travel, and leisure; Supreme Court data on names, birth year, and sex; National Health Insurance Service Health Checkup Information as of December 31, 2024; materials from the Korea Rural Economic Institute's 2024 Food Consumption Behavior Survey; and seed data and domain expertise contributed by NAVER Cloud during dataset design [15]. It does not provide a common reference year for every source. Consistent with synthetic data evaluation guidance, the dataset was evaluated as a potential auxiliary analytic layer, not as a record-level representation of real individuals or an official population estimate [12,13].

The CHS 2021-2025 was used as the main survey benchmark because CHS is designed to produce local community health statistics using a complex survey design [16,17]. The final 2025 CHS microdata were officially released on February 25, 2026 [17]. We included all 1,156,122 adult respondents with positive person weights and valid design identifiers: 229,242 in 2021, 231,785 in 2022, 231,752 in 2023, 231,728 in 2024, and 231,615 in 2025. CHS person weights were used to estimate weighted benchmark distributions [19]. The pooled design comprised 28,309 analytic strata and 112,299 analytic primary sampling units (PSUs). Survey year was appended to the CHS stratum and cluster identifiers. The unmet health care denominator was used only for analyses of the unmet health care indicator.

KOSIS public statistics–based indicators were used as seed-source–informed aggregate benchmarks because KOSIS provides official Korean statistical indicators across population and regional domains [18]. This analysis used 2 municipal-level indicators that were directly relevant to vulnerable-profile surveillance and available in the local analysis files: aged population share and older-adult living-alone household share. In this study, municipal-level areas correspond to Korean si/gun/gu units, including cities, counties, districts, and wards. The aged population indicator used 2025 KOSIS counts of persons aged 65 years or older divided by the total population. The older-adult living-alone indicator used the KOSIS older-adult one-person-household percentage; the local KOSIS export available for this indicator was 2024 and was carried forward into the 2025 comparison table. The KOSIS-persona comparison used 252 matched municipal-level areas: 216 exact si/gun/gu matches, 35 parent-city values repeated across administrative-gu rows where KOSIS reported a parent city, and one Sejong province-city match. Indicator definitions and linkage notes are provided in Table S2 in Multimedia Appendix 1. Because KOSIS denominators differ from persona adult rows and because KOSIS/public statistics were among the persona seed sources, KOSIS comparisons were interpreted as seed-source–informed regional ranking checks rather than fully independent absolute-level validation.

Variable Harmonization

Common core variables were harmonized across CHS and persona records: current province code, current municipal-level area code, age group, sex, marital status, living-alone status, and coarse education. These variables were selected because they are available in both the persona dataset and CHS and

XSL•FO
RenderX

correspond to common demographic and socioeconomic stratifiers used in health inequality monitoring [3,15-17]. Age was grouped into the following 5 categories: 19 to 29 years, 30 to 44 years, 45 to 64 years, 65 to 74 years, and 75 years or older. Education was classified into low, middle, and high categories. Economic activity and occupation were retained as sensitivity variables rather than core variables.

Economic activity and occupation required special handling. Persona economic activity was derived from generated occupation text, and persona occupation was a generated detailed job-title string, whereas CHS uses survey-coded economic activity and occupation variables [15-17]. We therefore excluded both economic activity and occupation from the core manuscript analysis and treated them as sensitivity variables. A separate occupation recoding quality assurance compared the original keyword mapping with an expanded sensitivity mapping.

Detailed variable mapping is provided in Table S1 in Multimedia Appendix 1. Briefly, region was harmonized to current province and municipal-level area codes, with historical code changes and renamed or reorganized administrative areas mapped to current codes through a reviewed crosswalk, age was grouped using identical cutoff points, sex and marital status were mapped to shared categories, and education was classified into 3 analytically comparable levels. CHS living-alone status was derived from household variables, and persona living-alone status was derived using a prespecified exact-match rule for the Korean-language family-type label denoting living alone. All 39 distinct family-type labels were audited, and the only ambiguous label was evaluated using alternative recoding rules. Missing or unmapped values were excluded only from comparisons involving the affected variable, and each distribution was normalized over its comparison-specific denominator. All 1 million persona rows were aged 19 to 99 years, so the adult restriction excluded no persona rows.

Benchmark Consistency Measures

For each benchmark comparison, we computed CHS weighted percentages and persona row percentages. Distributional consistency was summarized using percentage-point differences, standardized differences, total variation distance (TVD), and Jensen-Shannon divergence [20,21]. TVD was used as the main scalar summary because it quantifies the total amount of distributional mass that would need to be shifted for the 2 distributions to match [20]. For example, a TVD of 0.05 indicates that 5% of the distributional mass would need to be reassigned across categories for the CHS and persona distributions to match [20]. Prior synthetic data evaluation work has used TVD across univariate, bivariate, and 3-way empirical marginal distributions as a model-free measure of synthetic data fidelity, with holdout data used as a reference for sampling noise rather than a fixed universal cutoff [9,12,13]. Accordingly, we interpreted TVD descriptively and comparatively, focusing on whether discrepancies increased from marginal distributions to higher-dimensional profiles and whether they persisted after sparse-cell restrictions. We did not apply a universal TVD threshold to declare 2 empirical distributions as equivalent or nonequivalent [9,12,13].

CHS cell-level sampling uncertainty was estimated using a Taylor-linearized ratio estimator where the number of cells was computationally tractable, consistent with design-based analysis of complex health surveys [19]. For each benchmark cell, the weighted cell proportion was estimated as the ratio of the cell-specific weighted total to the total weighted denominator. The linearized residual for each analytic primary sampling unit was calculated as the cluster cell weight total minus the estimated cell proportion multiplied by the cluster total weight; stratum-level variance contributions were then summed across analytic strata and divided by the squared total weight. As a prespecified analytic handling rule, singleton strata were assigned the average variance contribution from nonsingleton strata for the same cell when needed. CIs were calculated on the percentage scale using the 95% normal critical value and truncated to the 0% to 100% range. Cell-level uncertainty estimates were not generated for very high-dimensional comparisons when the number of cell-by-stratum combinations exceeded the prespecified computational threshold of 8,000,000. Sparse cells were interpreted as diagnostic signals rather than definitive population truths.

To quantify sampling uncertainty in scalar distributional measures, we additionally used a 500-replicate stratified cluster bootstrap for the TVD and Jensen-Shannon divergence estimates reported in Tables 2 and 3 [22]. CHS analytic PSUs were resampled with replacement within analytic strata defined by survey year and kstrata; persona distributions were treated as fixed. The cell universe was fixed to the observed union of CHS and persona cells. We report the observed plug-in estimate, bootstrap mean, and 2.5th to 97.5th percentile interval separately. In sparse high-dimensional universes, resampled CHS data occupy fewer cells than the full sample, which can produce positive plug-in TVD bias. We did not replace the observed estimate with a bootstrap bias–corrected estimate because such corrections can overcorrect under sparse multinomial support; the observed full-sample TVD was retained as the descriptive point estimate. For the CHS split-half reference, PSUs were randomly divided into 2 approximately equal groups within each design stratum. Weighted distributions were estimated in each half and their TVD was calculated over 100 random splits; the median values and 2.5th to 97.5th percentiles were reported. Because each half used half the survey sample, this reference is a conservative upper bound for full-sample sampling noise.

XSL•FO
RenderX

Table 2. Marginal sociodemographic benchmark consistency between Korea Community Health Survey estimates and synthetic persona rows[a].

Variables	K[b]	TVD[c] (95% bootstrap percentile interval)	JS[d] divergence (95% bootstrap percentile interval)	Maximum absolute difference (percentage points)	Mean absolute difference (percentage points)
Age group	5	0.013 (0.012-0.014)	0.000 (0.000-0.000)	0.99	0.52
Sex	2	0.000 (0.000-0.001)	0.000 (0.000-0.000)	0.02	0.02
Province	17	0.006 (0.005-0.006)	0.000 (0.000-0.000)	0.27	0.07
Living-alone status	2	0.008 (0.007-0.009)	0.000 (0.000-0.000)	0.82	0.82
Education	3	0.059 (0.058-0.061)	0.002 (0.002-0.002)	5.94	3.96
Marital status	4	0.039 (0.038-0.040)	0.001 (0.001-0.001)	3.89	1.94

[a]Intervals are fixed-cell-universe percentile bootstrap intervals from 500 stratified cluster bootstrap replicates of Korea Community Health Survey analytic primary sampling units; persona distributions were treated as fixed. Economic activity and occupation were moved to sensitivity analyses. Values displayed as 0.000 are positive values that round off to 0.000 at the reported precision rather than exact zeros.

[b]K: number of union cells.

[c]TVD: total variation distance.

[d]JS: Jensen-Shannon.

Table 3. Core joint and high-dimensional sociodemographic benchmark consistency between Korea Community Health Survey and synthetic personas, excluding economic activity and occupation[a].

Combination	K[b]	Observed TVD[c]	Bootstrap mean[d] (95% bootstrap percentile interval)	CHS[e] split-half TVD[f]	Independence baseline TVD[g]
Age×sex	10	0.015	0.015 (0.014-0.016)	0.003 (0.002-0.005)	0.019
Age×living-alone status	10	0.020	0.021 (0.019-0.022)	0.003 (0.002-0.005)	0.037
Age×sex×living-alone status	20	0.023	0.023 (0.022-0.024)	0.004 (0.003-0.006)	0.044
Age×sex×education	30	0.063	0.063 (0.062-0.064)	0.005 (0.003-0.007)	0.258
Province×age×sex×living-alone status×education	1015	0.080	0.080 (0.079-0.081)	0.025 (0.023-0.028)	0.272
Municipal area×age×sex×living-alone status×education[h]	14,550	0.117	0.122 (0.120-0.123)	0.082 (0.080-0.084)	0.295
National core high-dimensional profile[i]	238	0.106	0.106 (0.105-0.107)	0.009 (0.008-0.011)	0.451

[a]Economic activity and occupation were moved to sensitivity analyses because these variables had greater harmonization uncertainty.

[b]K: number of union cells.

[c]TVD: total variation distance.

[d]The bootstrap column reports the replicate mean followed by the 2.5th-97.5th percentile interval.

[e]CHS: Korea Community Health Survey.

[f]The Korea Community Health Survey split-half total variation distance is the median with its 2.5th-97.5th percentile interval across 100 within-stratum random primary sampling unit splits.

[g]The independence baseline was constructed by multiplying Korea Community Health Survey marginal distributions.

[h]For the municipal profile, the bootstrap interval is above the observed total variation distance because bootstrap resampling increased sparsity; this comparison should be interpreted alongside the number of union cells and the Korea Community Health Survey split-half reference.

[i]The national core high-dimensional profile combines age group, sex, marital status, living-alone status, and education.

Analytic Strategy

We conducted 8 analyses. First, we compared marginal distributions for core harmonized variables. Second, we compared low-dimensional joint distributions, including age by sex, age by living-alone status, age by sex by living-alone status, and age by sex by education. Third, we evaluated core national, provincial, and municipal high-dimensional profiles excluding economic activity and occupation. Fourth, we compared policy-relevant vulnerable profiles that did not depend on economic activity or occupation recoding. Fifth, we performed sensitivity analyses for economic activity, occupation, occupation recoding, living-alone recoding, and economic activity–based vulnerable profiles. Sixth, we performed sparse-cell sensitivity analyses by restricting comparisons to cells with CHS sample sizes of at least 30, 50, or 100. Seventh, we compared persona regional profiles with KOSIS public statistics benchmarks using Pearson and Spearman correlations,

XSL•FO
RenderX

including exact-match and conditional living-alone sensitivity analyses [18]. Eighth, we conducted additional uncertainty and robustness analyses using CHS temporal-stability diagnostics; within-stratum PSU split-half TVD references; product-of-CHS-margins independence baselines; and national, provincial, and municipal composition-standardized unmet health care analyses. These analyses were used as descriptive references for benchmark variability and fitness for use, not as formal equivalence tests.

We interpreted these analyses in terms of practical surveillance readiness using evidence triangulation rather than fixed TVD cutoffs. A persona layer was interpreted as more suitable for descriptive planning when discrepancies were small in absolute percentage-point terms: TVD was low relative to lower-dimensional benchmark comparisons, persona-versus-pooled-CHS TVD did not exceed relevant CHS variability references, and KOSIS correlations supported regional ranking consistency. Policy-relevant vulnerable profiles were considered usable for scenario generation or dashboard prototyping only when benchmark differences and uncertainty diagnostics were displayed alongside persona estimates. A profile was interpreted as requiring calibration before inferential surveillance use when discrepancies increased with dimensionality, exceeded CHS variability references, or depended on variables with substantial harmonization uncertainty. Independence baselines were interpreted in relation to the strength of the underlying CHS dependence. Profiles whose discrepancy narrowed toward the CHS split-half reference after sparse-cell restriction were interpreted as weaker evidence of residual discrepancy than profiles that remained far above it.

Analysis Software and Reproducibility

All data processing and statistical analyses were conducted in Python (Python Software Foundation) using scripted workflows. CHS weighted percentages were calculated from person weights within each comparison-specific denominator, whereas persona percentages were calculated from synthetic row counts [19]. For aligned distributions, TVD and Jensen-Shannon divergence were calculated from normalized proportions, bootstrap intervals resampled CHS analytic PSUs within strata, sparse-cell restrictions used unweighted CHS cell counts, and KOSIS comparisons used Pearson and Spearman correlations across matched municipal-level units [18]. The analysis scripts are publicly archived on Zenodo [23].

Ethical Considerations

The Institutional Review Board of Sunchon National University (1040173-202606-HR-032-02) approved this study for an exemption from review. This study used deidentified CHS benchmark estimates obtained through official data-use procedures, public KOSIS aggregate indicators, and synthetic persona rows [15-18]. The analysis did not involve participant recruitment, intervention, direct contact with individuals, or identifiable private information. The synthetic persona records do not represent observed individuals, although synthetic data still require explicit evaluation of utility, privacy, and intended-use fitness before downstream use [8,12,13]. Informed consent was not applicable to this secondary analysis of deidentified, aggregate, and synthetic data.

Results

Data Sources and Harmonized Variables

The persona dataset, CHS, and KOSIS played distinct benchmark roles (Table 4). Personas were evaluated as synthetic individual rows. CHS provided the main weighted survey benchmark. KOSIS provided seed-source–informed public statistics benchmarks for regional aging-related vulnerability.

Table 4. Data sources, benchmark roles, analytic units, and key variables.

Data source	Role	Unit	Key variables	Benchmark interpretation
Nemotron-Personas-Korea	Synthetic persona dataset	Synthetic individual row (n=1,000,000; age 19-99 years)	Region, age, sex, marital status, and family type used to derive living-alone status, education, occupation	Validation target; not observed surveillance data
CHS[a] 2021-2025	Survey benchmark	Weighted adult survey respondent (n=1,156,122; 28,309 strata; 112,299 PSUs[b])	Common sociodemographic variables, survey weights and design variables; unmet health care need for downstream sensitivity analysis	Complex-sample survey benchmark, comparison-specific valid denominators, and not a population gold standard
KOSIS[c,d] public statistics	Aggregate benchmark	Municipal-level aggregate indicator	Aged-population share and older living-alone household share	Seed-source–informed consistency benchmark and denominators differ from persona rows

[a]CHS: Korea Community Health Survey.

[b]PSU: primary sampling unit.

[c]KOSIS: Korean Statistical Information Service.

[d]KOSIS was treated as a seed-source–informed public statistics benchmark rather than an independent validation source because KOSIS/public statistics were among the persona seed sources.

Marginal Benchmark Consistency

Basic demographic margins showed close agreement between persona rows and CHS survey benchmarks (Table 2). TVD was <0.001 (bootstrap percentile interval 0.000-0.001) for sex, 0.006 (bootstrap percentile interval 0.005-0.006) for province, 0.008 (0.007-0.009) for living-alone status, and 0.013 (bootstrap

XSL•FO
RenderX

percentile interval 0.012-0.014) for age group. Larger differences were observed for marital status (TVD 0.039, bootstrap percentile interval 0.038-0.040) and education (TVD 0.059, bootstrap percentile interval 0.058-0.061).

Economic activity and occupation were excluded from the core marginal table because both required interpretation of generated occupation text; their results are reported as sensitivity analyses in Table S5 in Multimedia Appendix 1.

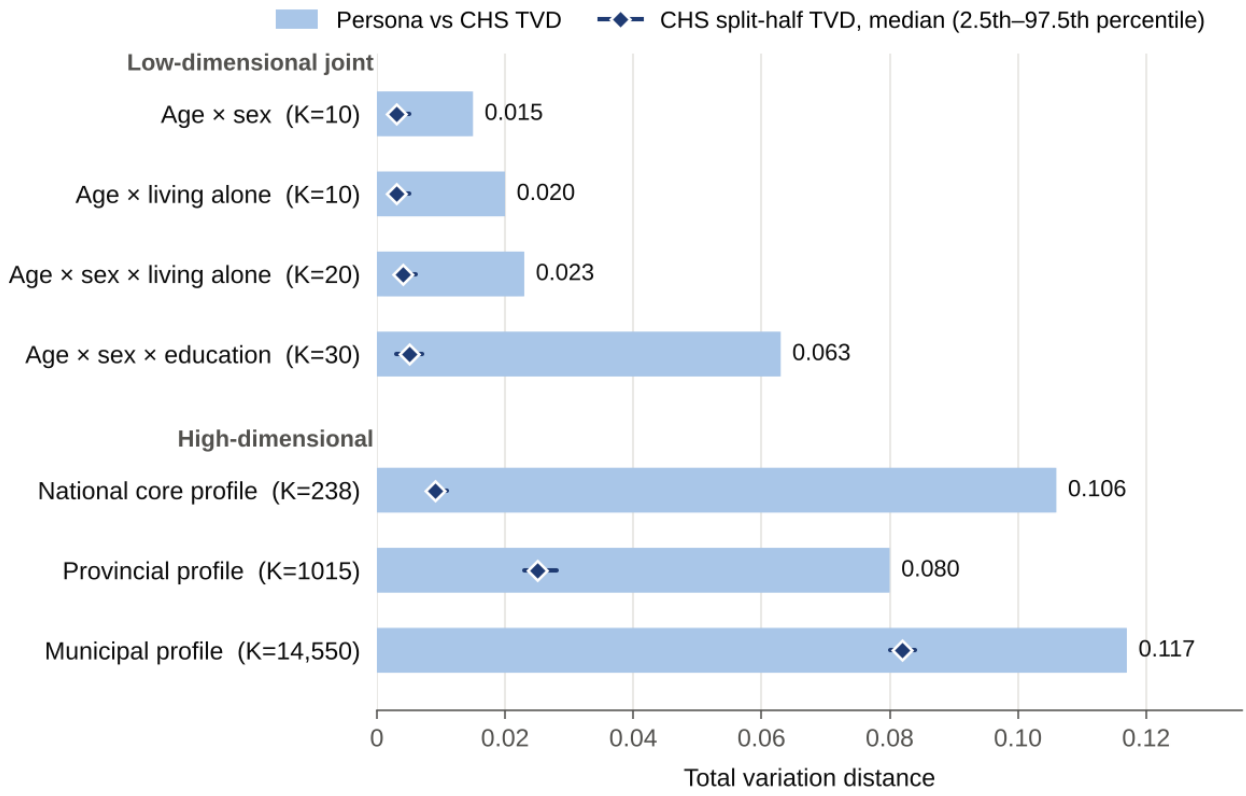
Joint Benchmark Consistency Excluding Occupation

Joint-distribution discrepancies increased with profile dimensionality (Table 3). TVD was 0.015 (bootstrap percentile interval 0.014-0.016) for age by sex, 0.020 (bootstrap percentile interval 0.019-0.022) for age by living-alone status, 0.023 (bootstrap percentile interval 0.022-0.024) for age by sex by living-alone status, and 0.063 (bootstrap percentile interval 0.062-0.064) for age by sex by education.

The largest core discrepancies appeared in higher-dimensional profiles. The provincial 5-way profile had observed TVD 0.080, a bootstrap mean of 0.080 (0.079-0.081), and a CHS split-half median of 0.025 (2.5th-97.5th percentile 0.023-0.028). The national 5-way profile had TVD 0.106 (bootstrap mean 0.106; 0.105-0.107), substantially above its split-half median of 0.009 (2.5th-97.5th percentile 0.008-0.011). The municipal 5-way profile had observed TVD 0.117, a bootstrap mean of 0.122 (0.120-0.123), and a split-half median of 0.082 (2.5th-97.5th percentile 0.080-0.084). The municipal bootstrap distribution lay above the full-sample estimate because resampling increased sparsity in the 14,550-cell universe. Relative to independence baselines, the gain was small for age by sex (0.015 vs 0.019), where CHS dependence was weak, but large for age by sex by education (0.063 vs 0.258), where dependence was stronger. Thus, the national discrepancy was clear, whereas the municipal comparison required interpretation alongside its sparse-cell and split-half diagnostics.

In Figure 1, the multiplication sign "x" denotes cross-classified joint distributions. The municipal-level profile comprises municipal area, age, sex, living-alone status, and education. The provincial profile comprises province, age, sex, living-alone status, and education. The national high-dimensional profile comprises age, sex, marital status, living-alone status, and education. Economic activity and occupation are excluded from this core figure and evaluated in sensitivity analyses. Bars show persona-versus-CHS TVD; diamonds and horizontal lines show the CHS split-half median and 2.5th to 97.5th percentile range across 100 within-stratum primary sampling unit splits. "K" denotes the number of union cells. Municipal-level areas correspond to Korean si/gun/gu units, including cities, counties, districts, and wards.

Figure 1. Total variation distance between the synthetic persona data and CHS benchmark across progressively richer sociodemographic profiles, with CHS split-half references.



Policy-Relevant Vulnerable Profiles

Policy-relevant vulnerable profiles showed moderate national-level agreement (Table 5). Adults aged 65 years or older who were widowed women represented 5.10% of the CHS weighted benchmark and 6.02% of persona rows, a difference of 0.93 percentage points. Adults aged 65 years or older with low education and living alone represented 3.59% of CHS and

XSL•FO
RenderX

3.30% of persona rows, a difference of –0.29 percentage points. Differences were –0.13 percentage points for women aged 75 years or older living alone and 0.01 percentage points for adults aged 19 to 29 years living alone. Economic activity–based vulnerable profiles were not treated as core findings and are reported separately in Table S6 in Multimedia Appendix 1.

In Table 5, profiles are cross-classified sociodemographic groups selected for public health surveillance relevance, including older adults, women, living-alone status, and low education. Economic activity–based vulnerable profiles are reported in Table S6 in Multimedia Appendix 1. Differences were calculated from unrounded percentages.

Table 5. Policy-relevant vulnerable profile consistency between CHS[a] and synthetic personas.

Vulnerable profile	CHS weighted %, survey benchmark	Persona %, synthetic rows	Difference, pp[b]	CHS 95% CI
Age 65+ years, widowed female	5.10	6.02	0.93	5.05-5.15
Age 65+ years, low education, living alone	3.59	3.30	–0.29	3.55-3.63
Female, 75+ years, living alone	2.07	1.94	–0.13	2.04-2.10
Age 19-29 years, living alone	2.68	2.69	0.01	2.63-2.74

[a]CHS: Korea Community Health Survey.

[b]pp: percentage point.

Occupation and Living-Alone Harmonization Sensitivity

Economic activity and occupation harmonization were evaluated in Table S5 in Multimedia Appendix 1, and vulnerable profiles depending on economic activity were reported in Table S6 in Multimedia Appendix 1. Economic activity had marginal TVD 0.017 (95% bootstrap percentile interval 0.016-0.018), and the profile adding economic activity had TVD 0.137 (95% bootstrap percentile interval 0.136-0.138). Occupation had marginal TVD 0.271 (95% bootstrap percentile interval 0.271-0.272), which decreased to 0.149 under the expanded keyword mapping; the occupation-containing high-dimensional profile had TVD 0.368 (95% bootstrap percentile interval 0.367-0.369). Economic activity and occupation were therefore retained as sensitivity evidence rather than core benchmark domains. Living-alone recoding sensitivity was evaluated separately in Table S9 in Multimedia Appendix 1. The living-alone results were robust to alternative treatment of the single ambiguous family-type label: marginal TVD ranged from 0.008 to 0.014, national core TVD from 0.105 to 0.106, and municipal core TVD was 0.117 under all 3 rules.

As a regional downstream compositional sensitivity analysis, we applied national CHS profile–specific medical-need and unmet-need probabilities to CHS and persona profile distributions (Table S7 in Multimedia Appendix 1). Using a primary minimum cell count of 30, the mean absolute difference between CHS- and persona-composition standardized rates was 0.076 percentage points across 17 provinces and 0.109 percentage points across 252 municipalities. Municipal standardized rates had Pearson r=0.567 and Spearman ρ=0.683 across the 2 compositions. However, composition-standardized rates did not reproduce observed municipal CHS unmet-need rates (persona composition: mean absolute difference 1.635 percentage points; Pearson r=0.004). Because CHS- and persona-composition rates differed by only 0.109 percentage points on average, CHS-composition rates also departed from observed municipal rates by at least 1.5 percentage points on average. Therefore, the lack of agreement with observed local rates reflects the limits of standardization with national profile-specific probabilities rather than persona composition alone. This analysis evaluated compositional sensitivity, not health outcome prediction.

Sparse-Cell and Sampling-Uncertainty Findings

Sparse-cell sensitivity differed materially between national and municipal profiles (Table S8 in Multimedia Appendix 1). The national core TVD remained 0.105 after restriction to CHS cells with n≥100, far above its split-half median of 0.009 (2.5th-97.5th percentile 0.008-0.011). In contrast, municipal TVD decreased from 0.117 across all cells to 0.105 at n≥30 and 0.091 at n≥100, with 73.0% of CHS mass and 69.2% of persona mass retained at n≥100. The all-cell municipal TVD exceeded its split-half median of 0.082 (2.5th-97.5th percentile 0.080-0.084), which is itself a conservative upper bound for full-sample sampling noise. Because the split-half reference was estimated across all cells rather than within each restricted cell universe, the restricted TVD values are not directly comparable with it. The municipal discrepancy was therefore smaller and less clearly separated from CHS sampling noise than the national discrepancy, but it was not explained by sparse cells alone. For the occupation-containing sensitivity profile, TVD decreased from 0.368 to 0.232 at n≥100; this restriction retained 96.9% of CHS mass but only 70.2% of persona mass, indicating that 29.8% of persona rows occupied combinations that were rare or absent in CHS.

CHS Temporal Stability and CI Diagnostics

The temporal-stability analysis compared persona-versus-pooled-CHS TVD with year-specific CHS-versus-pooled-CHS TVD for the same benchmark layers. For age group, the persona TVD was 0.013, below the maximum CHS year-versus-pooled TVD of 0.020. For living-alone status, the persona TVD was 0.008, also below the maximum CHS yearly TVD of 0.012. These results support the interpretation that the age and living-alone margins were close relative to ordinary CHS year-to-year variation.

XSL•FO
RenderX

In contrast, discrepancies for education and higher-dimensional profiles exceeded the CHS temporal-stability reference (Table S4 in Multimedia Appendix 1). Education had persona-versus-pooled CHS TVD 0.059 compared with maximum CHS year-versus-pooled TVD 0.016. Age by sex by education had persona TVD 0.063 compared with maximum yearly TVD 0.037. The national core profile had persona TVD 0.106 compared with maximum yearly TVD 0.045. The municipal profile had persona TVD 0.117 compared with maximum yearly TVD 0.091; this smaller difference should be interpreted with the municipal split-half and sparse-cell results.

CHS 95% CI diagnostics were intentionally interpreted cautiously because the pooled CHS benchmark included more than 1 million observations and therefore produced narrow CIs. For several margins and vulnerable profiles, persona percentages fell outside the pooled CHS 95% CIs despite absolute differences of less than 1 percentage point. For example, the maximum age-group difference was 0.99 percentage points. These diagnostics show that statistical distinguishability from a large survey benchmark does not by itself imply public health invalidity; substantive interpretation should consider absolute percentage-point differences, TVD, temporal CHS stability, sparse cells, and intended surveillance use.

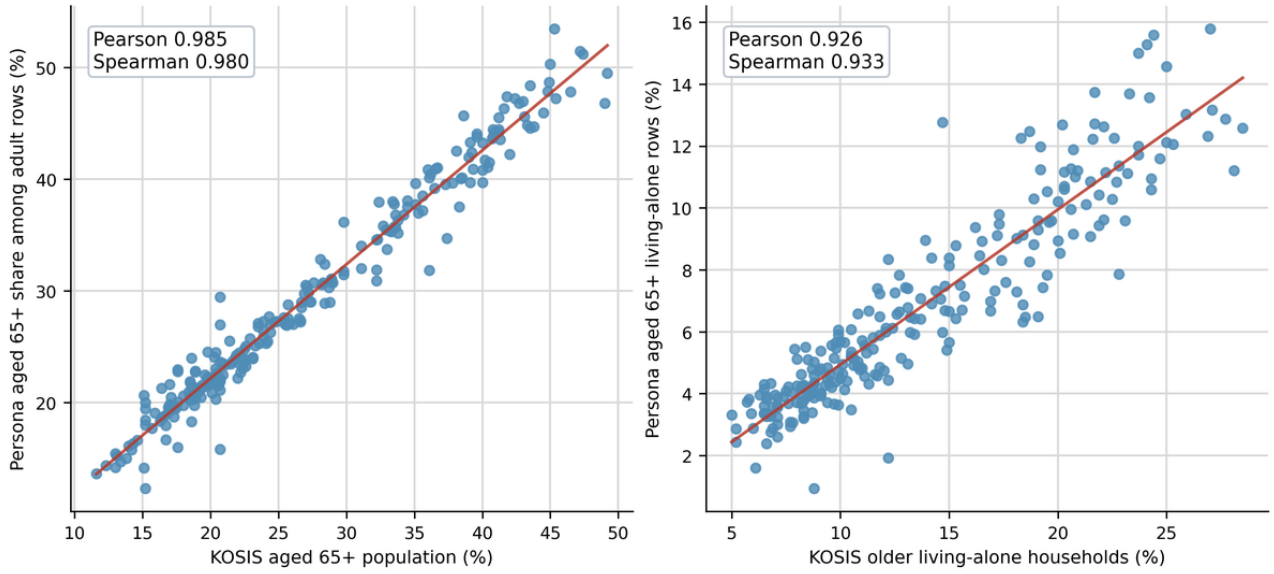
KOSIS Public Statistics Benchmark Consistency

KOSIS public statistics comparisons showed strong regional ranking consistency for age structure but weaker consistency for the stricter conditional living-alone comparison (Figure 2 and Table S3 in Multimedia Appendix 1). Across 252 regions, the correlation between KOSIS aged-population share and persona aged-adult-row share was 0.985 by Pearson correlation and 0.980 by Spearman correlation. The composite proxy comparing KOSIS older-adult living-alone household share with persona aged-65-plus living-alone adult-row share was Pearson correlation of 0.926 and Spearman correlation of 0.933. However, the stricter conditional comparison of persona living-alone share among rows aged 65 years or older with the KOSIS older-adult living-alone household indicator was weaker (Pearson r=0.547; Spearman ρ=0.576). Exact municipal-match sensitivity produced similar conclusions: aged-population ranking remained strong (Pearson r=0.989; Spearman ρ=0.991), whereas conditional older-adult living-alone consistency remained modest (Pearson r=0.530; Spearman ρ=0.550).

Because KOSIS indicators used total population or household denominators while personas represented adult synthetic rows, absolute-level differences were not interpreted as direct validation. The KOSIS results indicate that the persona data captured regional age-structure rankings well, but they provide weaker support for living-arrangement structure among older adults after conditioning on age.

In Figure 2, points represent matched municipal-level areas, corresponding to Korean si/gun/gu units; fitted lines show linear trends. The left panel compares the KOSIS percentage of residents aged 65 years or older with the corresponding persona share among adult rows. The right panel compares the KOSIS older-adult one-person-household percentage with the corresponding persona share of rows that were both aged 65 years or older and living alone among all adult rows. Because these KOSIS indicators differ from persona row denominators and because KOSIS was among the public statistical sources used to construct the persona data, this figure is interpreted as a seed-source–informed public statistics benchmark check rather than independent validation; exact-match and conditional living-alone sensitivity results are reported in Table S3 in Multimedia Appendix 1.

Figure 2. Municipal-level regional consistency between KOSIS public-statistics indicators and corresponding synthetic persona profiles. KOSIS: Korean Statistical Information Service.



Summary of Surveillance Readiness

The validation layers support a tiered interpretation for surveillance use. The persona data appear suitable for descriptive surveillance planning when the use case depends on basic demographic margins and regional age-structure rankings. Policy-relevant vulnerable profiles may be useful for scenario generation and dashboard prototyping if benchmark differences

XSL•FO
RenderX

and uncertainty diagnostics are reported alongside persona estimates. The national high-dimensional discrepancy indicates a need for calibration. At the municipal level, the discrepancy exceeded the all-cell CHS split-half reference but narrowed after sparse-cell restriction, so its separation from CHS sampling noise was less clear. Detailed local profiles should therefore be used as calibration diagnostics rather than direct evidence for resource allocation, subgroup inference, or local health outcomes.

Discussion

In this benchmark-consistency study, public statistics–based synthetic personas reproduced basic demographic margins and regional aging-related rankings but did not fully reproduce higher-dimensional individual profiles. Agreement was strongest for age, sex, province, and living-alone status; discrepancies became more apparent for education and profiles combining several sociodemographic characteristics. The national core profile remained clearly different from its CHS split-half reference after sparse-cell restriction, whereas the municipal profile exceeded its all-cell split-half reference by a smaller margin and its discrepancy narrowed after the sparsest cells were excluded. The study's main contribution is a practical distinction between row-level plausibility and readiness for surveillance use, consistent with synthetic data evaluation frameworks that distinguish resemblance, utility, privacy, and intended-use validity [9,12,13].

The observed pattern is plausible given how public statistics–based personas are constructed. Variables such as age, sex, and region were among the generator's public statistics inputs and can therefore be represented with relatively high fidelity [15]. Their agreement should be interpreted as consistency with design inputs rather than independent external validation. Regional aging-related indicators also have clear aggregate counterparts in KOSIS, which helps explain the strong regional ranking consistency observed in this study [18].

The larger discrepancies in education and high-dimensional profiles point to a different issue: preserving margins does not ensure preservation of dependence structures [10,11]. The independence baseline clarified this distinction. Age by sex showed little improvement over a product of CHS margins because the underlying dependence was weak, whereas age by sex by education was substantially closer to CHS than its independence baseline. Thus, the persona data retained some joint information but did not preserve enough of the joint structure to support uncalibrated subgroup inference.

This distinction is especially important for small-area surveillance. At the municipal level, CHS itself contains sparse high-dimensional cells, and its split-half reference is conservatively high because each half contains only half the sample [16,19,20]. The municipal persona TVD of 0.117 exceeded this conservative reference (median 0.082; 2.5th-97.5th percentile 0.080-0.084) and decreased to 0.091 after restriction to CHS cells with n≥100; because the split-half reference was estimated across all cells, the restricted value should not be read as falling within CHS sampling noise. A discrepancy in a sparse local profile is therefore better read as a signal for calibration or further review, rather than as a precise estimate of how far the persona distribution departs from the local population [12,13,19].

We did not base readiness on a single distributional cutoff. Instead, we combined absolute percentage-point differences, TVD patterns across increasing profile dimensionality, sparse-cell and CHS temporal-stability diagnostics, split-half and independence references, and KOSIS regional ranking consistency. The regional downstream analysis provided a separate fitness-for-use check: substituting persona for CHS demographic composition changed municipal standardized unmet-need rates by about 0.1 percentage points on average, although the municipal ranking of these rates was only moderately preserved (Pearson r=0.567). Neither composition reproduced observed municipal CHS rates, because standardization with national profile-specific probabilities does not capture local outcome variation. Composition-only analyses therefore do not establish that synthetic personas can reproduce local health outcomes.

This study extends prior work on synthetic populations by shifting attention from marginal plausibility to surveillance-relevant population structure. Synthea established the value of synthetic health records for public health informatics infrastructure, but its primary contribution was generation of realistic synthetic clinical histories and data standards rather than validation against small-area survey benchmarks [6]. The synthpop literature focused on producing synthetic versions of confidential microdata while balancing data utility and disclosure protection [7], and biomedical synthetic patient studies similarly evaluated synthetic data through utility and privacy lenses [8]. Population synthesis studies have long recognized that fitting target margins is not sufficient when downstream use depends on correlations among attributes. For example, Pritchard and Miller [11] addressed simultaneous fitting to household and person margins, and Jeong et al [10] used copulas to preserve dependence structures in synthetic population generation. Synthetic data fidelity frameworks have further evaluated empirical marginal and joint distributions, often interpreting distributional distances in relation to holdout data or sampling variability rather than universal thresholds [9]. In public health and epidemiologic applications, TVD has been used as a descriptive measure of distributional difference across age, sex, region, and other product spaces [24]. These precedents support TVD as a useful summary measure, but they do not justify a single cutoff for practical equivalence.

These findings show why this distinction matters for surveillance. The persona dataset performed well for several basic margins and for regional rankings of aging-related vulnerability. Departures became clearer when variables were combined into profiles closer to the way public health vulnerability is defined in practice. This pattern is consistent with the methodological concern that a synthetic population can look credible at the margin while losing important dependence structures.

The study also reinforces a broader principle in public health surveillance: no single data source should be treated as complete population truth. CHS provides observed survey information

XSL•FO
RenderX

and official small-area health indicators, but it remains a complex-sample survey with sampling uncertainty and sparse high-dimensional cells. KOSIS provides official aggregate statistics, but its denominators do not always match synthetic adult rows and KOSIS/public statistics were among the persona seed sources. In this setting, synthetic personas are best understood as a complementary analytic layer whose value depends on explicit validation and calibration, not as a replacement for survey or official statistics.

This interpretation is consistent with statistical quality frameworks that distinguish output quality, process quality, coherence, and fitness for intended use [25]. For a public statistics–based synthetic persona dataset, fitness for use should therefore be judged against the specific surveillance task: broad descriptive planning, profile construction, calibration, or inferential small-area estimation. These findings support the first 2 uses more strongly than the latter 2.

For public health informatics, the results support a cautious but constructive role for synthetic personas. They should not be used as stand-alone substitutes for observed CHS indicators, because CHS already provides observed small-area health measures and synthetic rows are not survey observations. A more defensible use is to support simulation, scenario analysis, vulnerability profiling, dashboard prototyping, and calibration workflows when benchmark differences, sparse-cell flags, and denominator caveats are visible to users.

A practical implication is that marginal agreement is not enough. Close alignment for age, sex, province, and living-alone status does not guarantee alignment for profiles such as older low-education adults living alone or for combinations of age, sex, marital status, education, and living arrangement. The exhaustive audit of all 39 family-type labels and the alternative recoding analysis supported retaining living-alone status in the core analysis, whereas occupation remained sensitive to keyword mapping.

For implementation, the results support a tiered approach to use. Basic demographic margins and regional aging-related rankings may be appropriate for descriptive planning and regional prioritization. Policy-relevant vulnerable profiles may support scenario generation if benchmark differences and uncertainty diagnostics are reported. The national high-dimensional profile requires calibration, whereas municipal profiles require both calibration and explicit consideration of CHS sampling noise before inferential use.

For future validations of public-statistics-based synthetic populations, basic margins are a necessary first check but not a sufficient one. Joint distributions are essential because they reveal discrepancies that marginal comparisons can hide. Policy-relevant vulnerable profiles should be reported alongside high-dimensional metrics because they are easier for surveillance users to interpret than thousands of sparse cells. Sparse-cell sensitivity should accompany CHS comparisons at detailed profile levels. Variables requiring substantial harmonization, such as occupation, should be separated into sensitivity analyses. Finally, distributional discrepancies should be interpreted against benchmark uncertainty and temporal stability where possible, rather than against arbitrary cutoffs.

The framework also points to where calibration should be directed. For variables with strong public statistics benchmarks, raking or poststratification could be used to align synthetic rows to selected CHS or KOSIS margins [26,27]. For surveillance tasks that depend on vulnerable profiles, iterative proportional fitting or related calibration methods could target selected joint distributions rather than all possible sparse cells [26,28]. This role is closer to formal small-area estimation and calibration workflows than direct use of synthetic rows as evidence; recent small-area estimation work illustrates how survey data and auxiliary population information can be combined to improve estimates for small domains [29]. Occupation requires a separate harmonization strategy, ideally validated against external labor statistics benchmarks before being used as a core surveillance variable. At the municipal level, calibrated outputs should retain sparse-cell flags and uncertainty annotations so that synthetic profiles are not overinterpreted as direct local population estimates.

The study used multiple benchmark layers, including CHS survey benchmarks, seed-source–informed KOSIS public statistics benchmarks, comparison-specific denominators, stratified cluster bootstrap intervals, sparse-cell and temporal-stability diagnostics, split-half and independence-baseline references, occupation and living-alone recoding checks, and regional downstream compositional sensitivity. This layered approach reduces the risk that conclusions are driven by a single benchmark or harmonization choice and makes the intended use of each result more explicit.

Several limitations remain. First, CHS is a complex-sample survey benchmark rather than a population gold standard. Second, CHS high-dimensional cells are sparse, especially at the municipal level; scalar TVD and Jensen-Shannon divergence intervals quantify CHS sampling variability but not uncertainty in the fixed persona generation, and repeated synthetic generations or model-based uncertainty estimates would be needed to evaluate that component. Third, KOSIS indicators and persona rows have different denominators, and KOSIS/public statistics were among the persona seed sources; KOSIS comparisons are therefore seed-source–informed consistency checks rather than independent validation. Fourth, the dataset card identifies 2024 reference information for some seed sources but not a common vintage for all inputs, so temporal mismatch with pooled 2021-2025 CHS cannot be fully resolved. Fifth, persona living-alone status remains a proxy derived from the structured family-type field, although all 39 labels were audited and the single ambiguous label did not materially affect the results (Table S9 in Multimedia Appendix 1); economic activity and occupation had greater harmonization uncertainty and were treated as sensitivity domains. Sixth, the regional downstream analysis standardized CHS outcome rates over alternative compositions and did not validate persona-generated health outcomes.

In this benchmark-consistency study, public statistics–based synthetic personas represented several basic demographic margins and regional aging-related patterns reasonably well but did not fully reproduce the high-dimensional profiles needed for subgroup surveillance. The national discrepancy was clear, whereas municipal evidence was limited by sparse cells and

XSL•FO
RenderX

CHS sampling noise. Synthetic personas may be useful for composition-based scenario development and dashboard prototyping but should not be treated as stand-alone sources for local surveillance or health-outcome inference without additional calibration and validation.

Acknowledgments

The author acknowledges the Korea Disease Control and Prevention Agency for providing access to Korea Community Health Survey data through official data-use procedures, the Korean Statistical Information Service (KOSIS) for providing publicly accessible statistical indicators, and NVIDIA for publicly releasing Nemotron-Personas-Korea through Hugging Face. The analyses and interpretations are solely those of the author and do not necessarily represent the views of the data providers or dataset publisher. During revision, the author used the generative AI tool Claude (Anthropic) to check consistency among the manuscript to suggest language edits. All statistical analyses were conducted by the author, who reviewed and verified all AI-assisted content and takes full responsibility for the accuracy and integrity of the manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT [MSIT]) (RS-2026-25591616) and by Korea Basic Science Institute(National research Facilities and Equipment Center) grant funded by the Korea government (MOE) (RS-2026-25533557).

Data Availability

The Korea Community Health Survey microdata used in this study are publicly accessible through the Korea Community Health Survey official website after following the survey's raw-data request and download procedures [17]. Korean Statistical Information Service aggregate indicators are publicly accessible through the Korean Statistical Information Service [18]. Nemotron-Personas-Korea is publicly accessible from its original Hugging Face repository [15]. The analysis scripts are publicly available on Zenodo [23].

Authors' Contributions

MK contributed to conceptualization, data curation, formal analysis, methodology, software, validation, visualization, writing the original draft, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional supporting data.
[DOCX File , 34 KB-Multimedia Appendix 1]

References

1. Shah SN, Russo ET, Earl TR, Kuo T. Measuring and monitoring progress toward health equity: local challenges for public health. Prev Chronic Dis. 2014;11:E159. [FREE Full text] [doi: 10.5888/pcd11.130440] [Medline: 25232746]
2. Pierannunzi C, Xu F, Wallace RC, Garvin W, Greenlund KJ, Bartoli W, et al. A methodological approach to small area estimation for the behavioral risk factor surveillance system. Prev Chronic Dis. 2016;13:E91. [FREE Full text] [doi: 10.5888/pcd13.150480] [Medline: 27418213]
3. World Health Organization. Handbook on Health Inequality Monitoring: With a Special Focus on Low- and Middle-Income Countries. Geneva. World Health Organization; 2013.
4. Rahman A. Estimating small area health-related characteristics of populations: a methodological review. Geospat Health. 2017;12(1):495. [FREE Full text] [doi: 10.4081/gh.2017.495] [Medline: 28555467]
5. Iyamu I, Xu AXT, Gómez-Ramírez O, Ablona A, Chang H, Mckee G, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill. 2021;7(11):e30399. [FREE Full text] [doi: 10.2196/30399] [Medline: 34842555]
6. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230-238. [FREE Full text] [doi: 10.1093/jamia/ocx079] [Medline: 29025144]
7. Nowok B, Raab GM, Dibben C. synthpop: bespoke creation of synthetic data in R. J. Stat. Soft. 2016;74(11):1-26. [doi: 10.18637/jss.v074.i11]
8. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. 2020;20(1):108. [FREE Full text] [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]

XSL•FO
RenderX

9. Platzer M, Reutterer T. Holdout-based empirical assessment of mixed-type synthetic data. Front Big Data. 2021;4:679939. [FREE Full text] [doi: 10.3389/fdata.2021.679939] [Medline: 34268491]
10. Jeong B, Lee W, Kim D, Shin H. Copula-based approach to synthetic population generation. PLoS One. 2016;11(8):e0159496. [FREE Full text] [doi: 10.1371/journal.pone.0159496] [Medline: 27490692]
11. Pritchard DR, Miller EJ. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation. 2012;39(3):685-704. [doi: 10.1007/s11116-011-9367-4]
12. Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. Methods Inf Med. 2023;62(S 01):e19-e38. [FREE Full text] [doi: 10.1055/s-0042-1760247] [Medline: 36623830]
13. Kaabachi B, Despraz J, Meurers T, Otte K, Halilovic M, Kulynych B, et al. A scoping review of privacy and utility metrics in medical synthetic data. NPJ Digit Med. 2025;8(1):60. [FREE Full text] [doi: 10.1038/s41746-024-01359-3] [Medline: 39870798]
14. El Emam K, Mosquera L, Fang X, El-Hussuna A. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. 2024;14(1):6978. [FREE Full text] [doi: 10.1038/s41598-024-57207-7] [Medline: 38521806]
15. NVIDIA. Nemotron-Personas-Korea. Hugging Face Datasets. URL: https://huggingface.co/datasets/nvidia/Nemotron-Personas-Korea [accessed 2026-05-14]
16. Kang YW, Ko YS, Kim YJ, Sung KM, Kim HJ, Choi HY, et al. Korea community health survey data profiles. Osong Public Health Res Perspect. 2015;6(3):211-217. [FREE Full text] [doi: 10.1016/j.phrp.2015.05.003] [Medline: 26430619]
17. Korea Disease Control and Prevention Agency. 2025 community health survey raw data user guide and reference materials [Korean]. Community Health Survey. URL: https://chs.kdca.go.kr/chs/mnl/mnlBoardMain.do [accessed 2026-09-30]
18. Korean Statistical Information Service (KOSIS). URL: https://kosis.kr/eng [accessed 2026-05-14]
19. Korn EL, Graubard BI. Analysis of Health Surveys. New York. Wiley; 1999.
20. Gibbs AL, Su FE. On choosing and bounding probability metrics. Int Statistical Rev. 2007;70(3):419-435. [doi: 10.1111/j.1751-5823.2002.tb00178.x]
21. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans. Inform. Theory. 1991;37(1):145-151. [doi: 10.1109/18.61115]
22. Rao JNK, Wu CFJ. Resampling inference with complex survey data. Journal of the American Statistical Association. 2012;83(401):231-241. [doi: 10.1080/01621459.1988.10478591]
23. Analysis code for the sociodemographic benchmark consistency study of Korean synthetic personas. Zenodo. URL: https://zenodo.org/records/23036268 [accessed 2026-10-20]
24. Schneckenreither G, Herrmann L, Reisenhofer R, Popper N, Grohs P. Assessing the heterogeneity in the transmission of infectious diseases from time series of epidemiological data. PLoS One. 2023;18(5):e0286012. [FREE Full text] [doi: 10.1371/journal.pone.0286012] [Medline: 37253038]
25. Gennari P, Smit J. Ensuring the quality of statistical processes and statistical outputs at the national level. Statistical Journal of the IAOS. 2025;41(3):499-507. [doi: 10.1177/18747655251368666]
26. Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. 1940;11(4):427-444. [doi: 10.1214/aoms/1177731829]
27. Deville JC, Särndal CE. Calibration estimators in survey sampling. J Am Stat Assoc. 2012;87(418):376-382. [doi: 10.1080/01621459.1992.10475217]
28. Beckman RJ, Baggerly KA, McKay MD. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice. 1996;30(6):415-429. [doi: 10.1016/0965-8564(96)00004-3]
29. Yilema SA. Bayesian small area estimation of anemia prevalence using a generalized linear mixed model. Stat J IAOS. 2025;41(1):171-179. [doi: 10.1177/18747655241308394]

Abbreviations

CC-BY: Creative Commons Attribution
CHS: Korea Community Health Survey
KOSIS: Korean Statistical Information Service
PSU: primary sampling unit
TVD: total variation distance

XSL•FO
RenderX

Edited by O Leal Neto; submitted 06.Jul.2026; peer-reviewed by J-Y Kim, H Hassani; comments to author 11.Aug.2026; revised version received 30.Sep.2026; accepted 01.Oct.2026; published 09.Nov.2026

Please cite as:
Kim M
Viability of Public Statistics–Based Synthetic Personas in Small-Area Public Health Surveillance: Sociodemographic Benchmark Consistency Study Using Korea Community Health Survey and Korean Statistical Information Service Data
JMIR Public Health Surveill 2026;12:e106306
URL: https://publichealth.jmir.org/2026/1/e106306
doi: 10.2196/106306
PMID:

©Miryoung Kim. Originally published in JMIR Public Health and Surveillance (https://publichealth.jmir.org), 09.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Public Health and Surveillance, is properly cited. The complete bibliographic information, a link to the original publication on https://publichealth.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX