Research article

Inferential Z-Test Validation of Dual Structural Bias in Cancer Risk Assessment within Large COVID-19 Vaccine Cohorts

Marco Roccetti [1]

[1] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

* **Correspondence:** Email: marco.roccetti@unibo.it; Tel: +393920271318

Abstract: The increasing reliance on complex analytical models using large administrative health data necessitates rigorous, pre-analysis inferential validation of methodological integrity, a principle often neglected in high-impact epidemiological studies. A recent large-scale cohort study reported statistical associations suggesting increased cancer risk following COVID-19 vaccination [1], findings that a prior statistical descriptive analysis had linked to a severe structural asymmetry, suggesting a pronounced external validity discrepancy [2]. Building upon the results of [2], this study aims to formally and inferentially validate this dual structural bias, specifically, the non-representativeness of the cohort's age distribution relative to the national population (Root Cause), and the resulting non-compatibility of the cancer incidence rate in the control subgroup relative to the national *gold standard* (External Validity). We applied two inferential tests, both Z-tests. The first tested for the demographic representativeness by comparing the sample proportion of individuals >= 65 years (12.2%) against the same national population proportion (18%). The second tested the incidence compatibility by comparing the non-vaccinated >= 65 subgroup Crude Incidence Rate against the national Crude Incidence Rate. The Z-test for demographic representativeness yielded a test statistic Z-score of - 260.39 (with a *p-value* $< 10^{-50}$), confirming a profound structural sampling failure (- 32.2%) already identified in [2]. The Z-test for incidence was equally conclusive, yielding a Z-score of - 15.23 (with again a *p-value* $< 10^{-50}$), formally validating an enormous cancer incidence deficit in that age bracket of the cohort. The combined inferential evidence confirms a dual structural bias caused by the systemic under-sampling of the high-risk elderly demographic. This non-equivalence means the baseline risk in the non-vaccinated group was artificially suppressed, causing mathematical inflation of the outcome and invalid results stemming from uncorrected baseline data. Our results mandate that the validity of conclusions drawn from any large-scale cohort study must be conditional upon the inferential confirmation of methodological integrity.

Keywords: Biomathematics, Computational Epidemiology; Inferential Statistics, Selection Bias,

COVID-19 vaccination, Cancer Research

1. Introduction

The current landscape of public health research is critically dependent on massive administrative health datasets, fueling the widespread adoption of complex epidemiological models. While this reliance promises faster and broader discoveries, it harbors a fundamental risk: the potential for methodological flaws or inherent selection biases present in the source data to be amplified, rather than mitigated, by the complexity of mathematical modeling. A core principle in biostatistics dictates that the rigor and clinical relevance of any complex model must be subordinate to, and contingent upon, the integrity and external validity of the underlying statistical design. We assert that complex models cannot compensate for fundamentally poor data structure. In situations where comparison groups are rendered non-equivalent or the entire cohort is non-representative, mathematical models, regressions for example, while performing mathematical control, can isolate and subsequently inflate the influence of this bias, leading to statistically significant conclusions that are clinically spurious.

A recent and highly influential large-scale retrospective cohort study, utilizing national South Korean administrative data, reported statistical associations suggesting an increased risk of cancer one year following COVID-19 vaccination [1]. Given the enormous clinical sensitivity and potential public health impact of such findings, the study's methodological foundation represents an exemplar case that need to be subjected to the highest level of inferential scrutiny. To this aim, we developed such a scrutiny with a preliminary mathematical analysis which identified two key structural asymmetries by simply comparing of the cohort of [1] against established national epidemiological and demographic gold standards [2, 3-7].

These preliminary descriptive observations established a causal chain of non-equivalence that this present study seeks to validate through formal inferential testing. This causal chain is based on two fundamental flaws. The first flaw, which can be defined as the demographic non-representativeness (Root Cause), was descriptively identified because the entire cohort possessed a proportion of elderly individuals >= 65 years that was substantially lower than the national average. Since age is the most powerful predictor of cancer, this imbalance suggested a structural sampling failure within the cohort itself. The second flaw, named as the risk non-equivalence (Consequence/External Validity Flaw), was descriptively observed as an 45.1% deficit in cancer incidence in the high-risk non-vaccinated control subgroup (>= 65 years) compared to the national gold standard [2]. This structural suppression of the baseline risk is the direct epidemiological result of the demographic flaw and the prerequisite for mathematically inflating the study's primary outcomes.

The present study's objective is instead to mathematically and inferentially validate these two descriptive findings using rigorous Z-tests. Our central hypothesis is dual, and causally linked: first, we hypothesize that the proportion of high-risk elderly individuals >= 65 years in the total cohort is statistically incompatible with the proportion of elderly individuals in the source national population (Root Cause confirmation); and second, we hypothesize that the deviation of the non-vaccinated group’s cancer incidence from the national rate is statistically significant to a level that rules out random chance, providing inferential proof that the baseline risk is artificially suppressed (Consequence confirmation). Inferential confirmation of this dual structural bias will provide conclusive evidence that the reported results of [1] suggesting harm are methodological artifacts derived from fundamental, uncorrected non-equivalence at baseline.

The remainder of this paper is organized as follows. The Materials and Methods Section details the application of two single-proportion Z-tests used to inferentially validate/reject the aforementioned representativeness and incidence. The Results section presents the obtained test statistics, followed by a Discussion Section that interprets the causal chain of the bias, its implications for results' validity, and limitations, culminating in a Conclusion Section that formally exposes the methodological artifact.

2. Materials and methods

2.1. Data Sources, Extracted Metrics and Descriptive Findings

Primary data, including Crude Incidence Rates (CRs), sample sizes (N), and age metrics, were extracted from the published paper by [1]. The National Gold Standard Benchmarks were sourced from official South Korean cancer statistics and demographic data [3-7]. Based on these national figures, the population aged >= years constitutes 18% of the total population in South Korea, and the crude incidence rate (CR) for the >= 65 population is 155.2 per 10,000 individuals. Instead, the total number of participants in the final matched study cohort of [1] was *N(Total)* = 2,975,035, as reported in Table 1 below.

Table 1. Input Parameters and National Benchmarks for Inferential Z-Tests.

Test	Parameter	Symbol	Value
Incidence (External Validity)	National Incidence (>= 65 years)	*P(0)*	155.2 per 10,000 (0.01552)
	Observed Incidence (>= 65 years, Non-Vacc.)	*p(inc)*	85.2 per 10,000 (0.00852)
	Non-Vaccinated Sample Size (>= 65)	*N(inc)*	72,285 participants
Age Representativeness	National Proportion (>= 65)	*P(age)*	18.0% (0.18)
	Total Cohort Size	*N(Total)*	2,975,035 participants
	Total Cohort >=65 Proportion	*p(age)*	12.2% (0.122)

2.2. Inferential Method I: Z-Test for the Age Representativeness

This test is uded to formally validate/reject the descriptive observation that the cohort's age

composition is non-representative. The test compares the observed proportion of individuals >= 65 years in the cohort *p(age),* that is *N(Cohort >= 65) / N(Total),* against the established national proportion *P(age)*.

The formal hypotheses can be consequently defined as follows:

> **Null Hypothesis (H0):** The cohort's proportion of individuals aged >= 65 is not statistically lower than the national proportion: *p(age) >= P(age)*;
>
> **Alternative Hypothesis (H1):** The cohort's proportion of individuals aged >= 65 is statistically lower than the national proportion (one-tailed test): *p(age) < P(age).*

Based on well known epidemiological formulas [8], the Z-Score derivation can be calculated as:

$$Z(Age\text{- }Prop) = \frac{p(age) - P(age)}{\sqrt{\frac{P(age)(1-P(age))}{N(Total)}}}, \quad (1)$$

2.3. Inferential Method II: Z-Test for the Cancer Incidence

This test formally verifies the hypothesis of a baseline cancer risk non-equivalence. It is carried out to have a confirmation/rejection of the hypothesis of consequence of the demographic issue on the cancer indicences. In essence, the test compares the observed cancer incidence rate in the non-vaccinated >= 65 subgroup *p(inc)* against the national gold standard rate *P(0)*. The formal hypothesis for this second test can be structured as follows:

> **Null Hypothesis (H0):** The observed incidence rate in the cohort's >= 65 subgroup is not statistically lower than the national rate: *p(inc) >= P(0)*;
>
> **Alternative Hypothesis (H1):** The observed incidence rate in the cohort's >= 65 subgroup is statistically lower than the national rate: *p(inc) < P(0).*

Consequently, the Z-Score derivation can be formulated as follows:

$$Z(Incidence) = \frac{p(inc) - P(0)}{\sqrt{\frac{P(0)(1-P(0))}{N(inc)}}}. \quad (2)$$

In closing these Sections, first it is to be noticed that both this latter test and the former one will be conducted as one-tailed tests, with a significance level of 0.05% to specifically verify the hypothesized deficits. Second we remind that the data presented here is either included directly or was extracted from the referenced documents. All calculations are easily reproducible based on the definitions provided. Further reasonable requests relative to data and calculations can be also addressed to the corresponding and sole author (email: marco.roccetti@unibo.it).

2.4. Ethics approval of research

This study uses publicly available, aggregated data that contains no private information. Therefore, ethical approval is not required.

3. Results

The combined results of the two inferential Z-tests earlier introduced provide overwhelming statistical evidence of a fatal dual structural bias as explained in the following two Sub-Sections.

3.1. Inferential Validation of Demographic Non-Representativeness

The formal validation of the demographic deficit was obtained by comparing the observed proportion of >= 65 individuals in the cohort *p(age)* against the national benchmark *P(age)*. All calculations and results are provided in details in Table 2 below.

Table 2. Calculation of Z-Score for Age Representativeness: *Z(Age - Prop)*.

Step	Description	Formula / Input Values	Calculated Value
I.A	Cohort Proportion *p(age)*	*N(Cohort >= 65)* / *N(Total)* = 361,425 / 2,975,035	0.122
I.B	National Proportion *P(age)*	National Benchmark	0.18
II	Numerator Calculation	*p(age) – P(age)*	0.122 - 0.18 = -0.058
III	Standard Error Calculation	$\sqrt{\frac{P(age)(1-P(age))}{N(Total)}}$	0.000223*
IV	Z-Score Calculation	II / III	-0.058 / 0.000223 = **-260.39**

To notice is the fact that for display purposes in the Table, the value 0.000223 is approximated; the real value is 0.00022274…, which is the one used in the Z-score calculation. In the end, as seen from the final row of Table 2, the resulting test statistic is: *Z(Age - Prop)* = - 260.39. The resultant *p-value* of $< 10^{-50}$ compels the categorical rejection of the Null Hypothesis of demographic compatibility. This Z-score value formally establishes that the entire study cohort is structurally non-representative of the source population, confirming the systemic relative deficit of -32.2% individuated in the previous preliminary analysis of [2] for the highest-risk demographic.

3.2. Inferential Validation of Cancer Incidence Flaw

The formal validation of the cancer incidence deficit (calculated as large as 45.1% in [2]) was obtained by comparing the observed rate in the non-vaccinated >= 65 subgroup *p(inc)* against the national gold standard (*P(0)*), using Formula 2 above, an yielding the results exposed in Table 3 below.

Table 3. Calculation of Z-Score for Cancer Incidence: Inferential verification of baseline risk non-equivalence against the national standard.

Step	Description	Formula / Input Values	Calculated Value
I.A	Observed Incidence *p(inc)*	85.2 per 10,000	85.5 / 10,000 = 0.00852
I.B	National Incidence *P(0)*	155.2 per 10,000	155.2 / 10,000 = 0.01552
II	Numerator Calculation	*p(inch) – P(0)*	0.00852 - 0.01552 = - 0.00700
III	Standard Error Calculation	$\sqrt{\frac{P(0)(1-P(0))}{N(inc)}}$	0.000460
IV	Z-Score Calculation	II / III	- 15.23

The resulting test statistic is: - 15.23. This Z-score is of an unprecedented magnitude, corresponding to a *p-value* that is statistically negligible ($< 10^{-50}$). The Null Hypothesis is conclusively rejected, formally establishing that the -45.1% deficit in cancer incidence, identified in the preliminary analysis [2], is a structural non-equivalence of risk.

In closing this Section, it should be noticed that our inferential results establish a statistically confirmed causal chain of bias. The profound demographic non-representativeness of the cohort *Z(Age - Prop)* = - 260.39, due to the systemic exclusion of the elderly population, creates a baseline population that is structurally too young. Since older age is the primary risk factor for cancer, this structural issue leads directly and inevitably to the artificially suppressed cancer incidence of - 15.41 in the reference non-vaccinated group, thereby creating the prerequisite mathematical condition for invalid results of [1].

4. Discussion

The inferential verification of the demographic non-representativeness, demonstrated by the Z-score of - 260.39, constitutes the most fundamental finding of this analysis. It formally confirms that the cohort studied in [1] is not a valid representation of the source Korean population, as it systemically excludes the high-risk elderly demographic (>= 65 years) at a statistically impossible magnitude. This demographic issue is the root cause of the study's overall structural bias and alone renders the entire study non-generalizable.

The second inferential result, Z-score of - 15.23, confirms the inevitable consequence of the demographic issue. The magnitude of this Z-score demonstrates that the 45.1% shortfall in cancer incidence in the non-vaccinated group is an epidemiological artifact of the non-representative sampling. The link is causal and inferentially verified: the cohort's statistical non-representativeness created a study population with an artificially suppressed baseline risk. This suppressed risk serves as the non-equivalent denominator in the calculation of the Hazard Ratio (HR) of [1], following the relationship: HR = Risk in Vaccinated Group / Risk in Non Vaccinated Group (Artificially Suppressed).

Consequently, the reported Hazard Ratios of [1], which suggest an increased risk of cancer (HR > 1), are methodological artifacts, that is mathematical amplifications of the pre-existing sampling bias, not biological effects of the vaccination. The dual evidence of non-representativeness and suppressed incidence (Z-scores of - 260.39 and - 15.23) provides irrefutable proof that the mathematical multivariable models employed on this data in [1] (i.e., Cox models) were rendered impotent by the structural bias in the input data. The model's failure is two-fold: an uncorrectable baseline (the outcome variable was fundamentally erroneous at baseline) and a structural confounding (the scale of the demographic non-representativeness is beyond the corrective capacity of standard modeling).

The unique and primary limitation of our present analysis rests on the necessity to use external national gold standards rather than having access to the individual patient-level data. However, the magnitude of the Z-scores obtained for both the demographic and the incidence issues is so extreme that it robustly overrides the influence of any reasonable unobserved confounders or minor deviations from the national gold standard, confirming a structural defect.

5. Conclusions

Our present study provides decisive inferential statistical evidence that the cohort data analyzed is an exemplar case of sampling from data that suffer from a fatal dual structural bias: a non-representativeness of the overall cohort relative to the source population (Z-score of - 260.39) that leads directly to a statistically incompatible cancer incidence rate in the control group (Z-score of - 15.23). We conclude that the study's reported Hazard Ratios [1] are overwhelmingly likely to be methodological artifacts resulting from an uncorrected structural bias inherent in the cohort selection process. This research mandates that the validity of conclusions drawn from any large-scale cohort study must be conditional upon the inferential confirmation of methodological integrity against both demographic and epidemiological *gold standards* before any serious epidemiological conclusion can be drawn [9-11].

Use of AI tools declaration

The author declares that he has not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This research received no external funding. The Author is grateful to several colleagues from the University of Bologna who provided comments on an earlier version of this paper distributed as a preprint.

Author's Contribution

MR conceived and designed the study, carried out all data collection and analysis, interpreted the quantitative results, and was the sole author responsible for writing and revising the manuscript. The author a^irms full responsibility for the integrity of the data and the accuracy of the data analysis presented.

Informed Consent Statement

Not applicable: Neither humans nor animals nor personal data are involved in this study.

Conflict of interest

The author declares there is no conflict of interest.

References

1. Kim HJ, Kim M-H, Choi MG, Chun EM. (2025) 1-year risks of cancers associated with COVID-19 vaccination: a large population-based cohort study in South Korea. *Biomark Res*. 13(114). DOI: 10.1186/s40364-025-00831-w

2. Roccetti M. (2025) A Biostatistical Reappraisal Unveiling the Mechanism Behind Apparent Cancer Risk Signals in a COVID-19 Vaccinated Cohort. *Zenodo Preprint* n. 17508347. DOI: 10.5281/zenodo.17508346

3. Kang MJ, Jung K-W, Bang SH, et al. (2023). Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2020. *Cancer Res Treat*., 55(2):385-399. DOI: 10.4143/crt.2023.447

4. Park EH, Jung K-W, Park NJ, et al. (2024). Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2021. *Cancer Res Treat*., 56(2):357-371. DOI: 10.4143/crt.2024.253

5. Park EH, Jung K-W, Park NJ, et al. (2025) Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2022. *Cancer Res Treat*. 57(2):312-330. DOI: 10.4143/crt.2025.264

6. Statista. South Korea: Cancer crude incidence rate by age, 2022. Statista; 2024. [Accessed 2025 Nov 2]. Available from: https://www.statista.com/statistics/1440818/south-korea-cancer-crude-incidence-rate-by-age/

7. World Bank. Population ages 65 and above (% of total population) - Korea, Rep. World Population Prospects, United Nations (UN). [Accessed 2025 Nov 2]. Available from: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?locations=KR 5

8. Rothman KJ, Greenland S, Lash TL. (2008) Measures of Disease Occurrence. In: *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. ISBN: 9780781755641

9. Roccetti, M., Cacciapuoti, G. (2025) Beyond the Gold Standard: Linear Regression and Poisson GLM Yield Identical Mortality Trends and Deaths Counts for COVID-19 in Italy: 2021–2025. *Computation* 2025, 13(10), 233. doi: 10.3390/ computation13100233

10. Chemaitelly H, Ayoub H, Coyle P, et al. (2025) Assessing healthy vaccinee effect in COVID-19 vaccine effectiveness studies: a national cohort study in Qatar. *eLife* 2025; 14:e103690. DOI: 10.7554/eLife.103690

11. Fisher RA, (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. A.*; 222:594-604. DOI: 10.1098/rsta.1922.0009