Research Article

Methodological Considerations on the External Validity of the Kim HJ et al. COVID-19 Vaccination Study (Biomark Res 13:114, 2025): A Quantitative Analysis

Marco Roccetti [1,*]

[1] Department of Computer Science and Engineering, University of Bologna, Italy; marco.roccetti@unibo.it

* Correspondence: marco.roccetti@unibo.it

Abstract

Background/Objectives: The Kim HJ et al. (2025) cohort study reported a surprising finding: a significantly higher 1-year cancer incidence risk in the COVID-19 vaccinated group. Given the vast public health implications, this result requires immediate and rigorous examination. This analysis provides the first formal external critique, integrating computational reviews to evaluate the methodology and test the external validity of the cohort.

Methods: Aggregated data from the Kim HJ et al. matched cohort (n=2,975,035) were used to calculate the overall Crude Incidence Rate (CR) of cancer. This study tested the resulting Crude Incidence Rate (CR) of the cohort against the established official national average CR for South Korea for the reference period (2020–2022). A secondary analysis of the final cohort's size ratio was performed to hypothesize a potential flaw in the 1:4 Propensity Score Matching (PSM) procedure.

Results: The cohort's overall CR was found to be 40.78 per 10,000, representing a substantial 22.26% downward deviation from the national average (52.46 per 10,000; SD 2.97). This discrepancy establishes a pronounced epidemiological paradox, strongly suggesting a lack of external validity for the cohort. Based on the exact 4:1 ratio of the final matched groups, the analysis proposes that the PSM procedure was likely inverted or misapplied, with the smaller unvaccinated group being used as the base cohort '1' for matching against the vaccinated group '4', which may have contributed to the suppression of the overall CR.

Conclusions: The reliability of the statistical associations reported by Kim HJ et al. is challenged by a possible lack of external validity and the hypothesized methodological ambiguities concerning the PSM. We conclude that independent validation is

mandatory and reiterate the call for public access to the underlying Korean National Health Insurance database to resolve these contradictions.

Keywords: COVID-19 Vaccination, Crude Incidence Rate, Cancer Incidence, Epidemiological Paradox, External validity , Propensity Score Matching, Inversion or Misapplication of the 1:4 Matching, Cohort Representativeness

1. Introduction

The accurate assessment of post-marketing adverse events, particularly those associated with widespread public health interventions like COVID-19 vaccination, is critical for public trust and effective health policy [1]. Observational studies drawing from national health insurance databases are essential tools in this process, offering large sample sizes and real-world data. However, the reliability of such studies is fundamentally dependent on the methodological rigor applied to cohort selection and statistical adjustment [2].

A core standard of rigor in epidemiology is External Validity. This concept refers to the extent to which the findings of a study can be generalized to other populations, settings, and circumstances outside the study's specific cohort. For a cohort derived from a national registry, high external validity requires that the study's overall burden of disease (measured by the Crude Incidence Rate, or CR) is statistically consistent with the known national burden of disease for the same period. Failure to meet this standard, often due to sampling or selection issues, means the cohort is not representative of the broader population, rendering its conclusions questionable in a real-world context.

The study by Kim HJ et al., published in *Biomarkers Research* in 2025 [3], is a retrospective, population-based cohort analysis utilizing data from some Korean National Health Insurance database to investigate the 1-year risks of cancers associated with COVID-19 vaccination in South Korea [4]. The study's finding, suggesting a higher rate of new cancer cases among the vaccinated population compared to the unvaccinated, is a surprising and scientifically challenging result that has yet to be fully addressed and scientifically analyzed with the required depth and urgency, especially considering the global scope of the vaccination programs [5].

The present work serves as a comprehensive critique that integrates two sequential analyses. Initially, a pronounced epidemiological paradox was identified based on raw incidence calculations derived from the study's supplementary data. This paradox established a significant *external inconsistency* between the study cohort's aggregate cancer incidence and official national statistics [6]. The follow-up analysis we developed [7] is as an integral part of the overall argument and posits a specific

methodological explanation for this paradox: the likely misapplication or inversion of the 1:4 Propensity Score Matching (PSM) procedure used by Kim HJ et al in [3].

In closing, the primary objective of this paper is to quantify and demonstrate the severity of the external inconsistency observed in the Kim HJ et al. cohort, thereby challenging its external validity. We then propose a plausible methodological hypothesis, specifically the inverted Propensity Score Matching (PSM), that would reconcile this numerical discrepancy and ultimately calls into question the reliability of the study's final association results. It is essential to undertake this critical examination using the known data, as the magnitude of the finding demands the highest level of scientific scrutiny.

In the remainder of this paper, Section 2 details the materials and methods for calculating the Crude Incidence Rates and formulating the PSM inversion hypothesis. Section 3 presents the quantitative results, including the derivation of the 22% deviation. Finally, Section 4 discusses the resulting lack of external validity, explores the PSM inversion as a probable cause of the bias, and provides conclusions and recommendations for data transparency.

2. Materials and Methods

In this Section, we provide all the necessary details on the data and methods used in this study, allowing readers to easily replicate our findings.

2.1. Sources of Data

This analysis is based entirely on publicly available, aggregated data extracted from the Kim HJ et al. study [3] and official South Korean national health statistics as reported in [8-10]. In particular, the raw cohort figures necessary for calculation were obtained from Table S4 ("Cumulative incidences of overall cancers in the matched cohort between vaccinated and unvaccinated individuals") in the Supplementary Material of the Kim HJ et al. manuscript [3]. These figures are as described in the following Table 1.

Table 1. Aggregated raw data from the Kim HJ et al. study, extracted from Table S4 of the supplementary materials [3], showing the final matched cohort counts and the reported Propensity Score Matching (PSM) ratio.

Metric	Value
Initial Cohort Size	8,407,849 individuals
Final Matched Cohort Size	8,407,849 individuals
Total Cancer Cases in Matched Cohort	12,133 cancer cases
Unvaccinated Group (N)	595,007 individuals
Unvaccinated Group (Cases)	1,989 cancer cases
Vaccinated Group (N)	2,380,028 individuals
Vaccinated Group (Cases)	10,144 cancer cases
Propensity Score Matching (PSM)	1:4 Ratio

The Official National Cancer Statistical data, including the Official Crude Incidence Rate (CR), for all cancers in South Korea were instead sourced from the Korean Central Cancer Registry for the years immediately preceding and during the study period (2020–2022), as reported in [8-10]. This data provides the robust national baseline against which the study cohort's representativeness is tested.

2.2. Definition and Calculation of Crude Incidence Rate (CR)

The Crude Incidence Rate (CR) per 10,000 population is a fundamental epidemiological measure used here specifically to evaluate the external validity of the cohort. Unlike Age-Standardized Rates (ASRs) which adjust for age distribution to allow comparison between populations, the CR reflects the raw burden of disease in a defined population over time [11]. Critically, any cohort derived from a national database should possess an aggregate CR that is statistically consistent with the national average CR for the same time period. A significant deviation signals a foundational problem in the initial sampling or selection process that any given procedure used to construct the cohort (like PSM for example) would fail to correct. The CR is calculated using the established epidemiological formula:

CR per 10,000 = (Number of new cases during a given period) / (Average population at risk during the same period) x 10,000 (1)

This is followed by a straightforward calculation of official South Korean CR baseline [8-10], whose values for both sexes per 100,000 population were converted to a per 10,000 basis to establish the national benchmark as described in Table 2 below.

Table 2. Official National Crude Incidence Rates (CR) for All Cancers in South Korea per 100,000 and the derived CR per 10,000, used to establish the national average baseline for the reference period (2020–2022).

Year	CR per 100,000	CR per 10,000
2020	482.9	48.29
2021	540.6	54.06
2022	550.2	55.02

Consequently, the official average CR baseline for all cancers for the reference period (2020–2022) can be established as the mean of these values: CR (Official Average) = 52.46 per 10,000 (Standard Deviation SD = 2.97). Finally, using the raw figures from Table S4 in the Supplementary material provided by Kim HJ et al. in [3], the following CRs of Table 3 are calculated using Eq. (1) for the cohort of interest.

Table 3. Calculated Crude Incidence Rates (CR) for the Kim HJ et al. matched cohort, showing the overall rate for the entire cohort and the rates for the segregated vaccinated and unvaccinated groups.

Group	Calculation (New Cancer Cases / Population) x 10,000	Crude Incidence Rate (CR)
CR (Cohort Overall)	(12,133 / 2,975,035) x 10,000	40.78 per 10,000
CR (Vaccinated)	(10,144 / 2,380,028) x 10,000	42.63 per 10,000
CR (Unvaccinated)	(1,989 / 595,007) x 10,000	33.43 per 10,000

2.3. Hypothesis Formulation on Propensity Score Matching (PSM) Inversion

A Propensity Score Matching (1:4 PSM) procedure aims to match each individual in the Treatment Group with four comparable individuals from the Control Group. In general, The Propensity Score Matching (PSM) is a quasi-experimental statistical method used to reduce the confounding bias that occurs when estimating the effect of a treatment or intervention (like vaccination in our case) in observational studies. The Propensity Score is defined as the conditional probability of an individual receiving the treatment given a set of observed covariates (e.g., age, sex, comorbidities). Hence, the propensity score $e(X)$ is given by $e(X)) = Prob\ (Z = 1 \mid X)$, where Z is the treatment assignment and X is the vector of baseline covariates.

The PSM calculation procedure involves a multi-step process. First, a logistic regression model is constructed to estimate the propensity score for every individual, based on the set of observed confounders. Once the propensity scores are calculated, the matching phase begins. Different matching algorithms exist (e.g., nearest neighbor, caliper, or kernel matching). In the reported 1:4 PSM, each treated individual (or the base group) is paired with four comparable control individuals whose propensity scores are nearly identical. This process effectively creates a synthetic, balanced cohort where the two groups are comparable on all measured confounders, thereby minimizing selection bias.

The primary rationale for using PSM is to mimic the randomization process of a randomized controlled trial in non-randomized observational data. By balancing the distribution of baseline covariates between the treated and control groups, PSM aims to isolate the true effect of the treatment (e.g., COVID-19 vaccination) from the effects of confounding factors that influenced the decision to vaccinate. If the PSM is successfully implemented, any residual difference in outcome between the matched groups can be more confidently attributed to the treatment itself. A failure in the PSM process, or a misapplication like the hypothesized inversion, fundamentally undermines this rationale and reintroduces significant bias into the analysis.

All this said, given the study's focus on the COVID-19 vaccine of [3], the standard and appropriate group assignment should have been: Treatment = Vaccinated and Control = Unvaccinated. The hypothesis of Inverted PSM can be hypothesized based on final reported cohort sizes: 595,007 Unvaccinated and 2,380,028 Vaccinated. This distribution is suggesting a reverse assignment: Base Group = Unvaccinated; Matched Group = Vaccinated.

3. Results

This Results Section presents two types of results: first, the quantification of the Epidemiological Paradox through the comparison of the calculated Crude Incidence Rate (CR) against the national baseline; and second, the numerical evidence supporting the hypothesis of Propensity Score Matching (PSM) inversion.

3.1. Quantification of the Epidemiological Paradox

The comparison between the study cohort's aggregated CR and the national average CR revealed a substantial and significant downward deviation, confirming the epidemiological paradox as summarized in the following Table 4.

Table 4. Quantification of the Epidemiological Paradox: Comparison of the Kim HJ et al. Cohort's overall Crude Incidence Rate (CR) against the Official National Average CR, highlighting the severe downward deviation

Metric	Rate per 10,000	Analysis
Official National Average CR (2020–2022)	52.46	Baseline for External Validity
Kim HJ et al. Cohort Overall CR	40.78	Calculated from Study Data
Downward Discrepancy	11.68	(52.46 - 40.78)
Percentage Deviation	≈ 22.26%	(52.46 - 40.78) / 52.46 x 10%

We have now the paradox summarized as follows: the study's analysis suggests an elevated cancer risk within the majority group (vaccinated CR is 27.5% higher than unvaccinated CR), which should intuitively push the overall cohort CR higher, yet the overall CR is 22.26% lower than the national baseline. This profound inconsistency represents a strong presumption of the cohort's lack of representativeness, which awaits formal refutation, though such a refutation appears mathematically challenging. To comprehend the full impact of this deviation, one must translate these statistical discrepancies into absolute numbers, which reveal the magnitude of the effect. Based on the cohort's overall rate of 40.78 per 10,000 and applying this to South Korea's population (approx. 51.77 million inhabitants [12]), the cohort rate would translate to approximately 211,273 new annual cancer cases. This is over 60,000 fewer new cases

than the 271,957 derived from the official national average rate of 52.46 per 10,000 for the same population. This massive deficit in expected cases underscores the profound lack of representativeness.

3.2. Evidence Supporting the PSM Inversion Hypothesis

The hypothesis that the 1:4 PSM was inverted is strongly supported by the final cohort sizes reported in Kim HJ et al.

In fact, given the study's focus on the COVID-19 vaccine, the standard and appropriate group assignment should have been: Treatment = Vaccinated; Control = Unvaccinated. The hypothesis of Inverted PSM is here formulated by analyzing the final reported cohort sizes: 595,007 Unvaccinated and 2,380,028 Vaccinated. This distribution is mathematically consistent with taking the smaller group (approx. 600,000) as the base "1" and matching it to the larger group (approx. 2.4 million) as the "4" component, suggesting the reverse assignment: Base Group = Unvaccinated; Matched Group = Vaccinated.

All this is further evidenced by the "numerical signature" where the total matched cohort (2,975,035) is exactly five times the size of the smaller unvaccinated group (595,007), that is the sum of the 1:4 ratio. This result confirms that the smaller unvaccinated group was used as the base '1' for the matching, thus inverting the standard procedure and yielding a final 4:1 ratio which is (erroneously) as follows: Ratio (Vaccinated / Unvaccinated) = 2,380,028 / 595,007 $\approx$ 4.00004.

This calculated ratio confirms the numerical correspondence: the vaccinated group (2,380,028) is precisely four times the size of the unvaccinated group (595,007). This exact numerical construction solidifies the argument for an inversion, where the small, unvaccinated pool defined the base cohort size for the 1:4 matching.

4. Discussion

The subsequent Discussion synthesizes the quantitative findings, focusing on three critical areas: 1) summarizing the previous background and the most relevant findings of this paper; 2) establishing the core Lack of External Validity stemming from the 22% downward CR deviation; 3) detailing the PSM Inversion Hypothesis as the primary methodological explanation for this bias; 4) contrasting the cohort against the South Korean Epidemiological Standards and 5) the limitations sorrounding this present study. Finally, Section 5 will conclude this treatment with recommendations for data transparency.

4.1. Summary of Prior Background and Key Findings

This research has integrated integrates two distinct lines of computational epidemiological analysis to critically evaluate the methodology and findings of the Kim HJ et al. (2025) cohort study concerning the 1-year risks of cancers associated with COVID-19 vaccination in South Korea. The Kim HJ et al. finding, suggesting a higher cancer incidence in the vaccinated group, is a surprising and scientifically challenging result that warrants rigorous, immediate scrutiny using all available data, as undertaken here. Our initial critique established a pronounced epidemiological paradox: while the Kim HJ et al. cohort suggested a higher crude cancer incidence rate (CR) among the vaccinated group compared to the unvaccinated group, the overall cohort CR was found to deviate downwards by over 22% from the official national average CR for South Korea recorded in the immediately preceding years (2020–2022). This fundamental discrepancy has suggested a lack of external validity for the cohort and raised strong concerns regarding its lack of representativeness.

Building upon this, the ultimate analysis has proposed a plausible explanation for the paradox: the reported 1:4 Propensity Score Matching (PSM) procedure was potentially inverted or misapplied. This inversion could introduce unidentified confounding factors and artificially bias the overall CR downwards. While we do not intend to reject the association results reported by Kim HJ et al. until direct access to the database is granted, the observed discrepancies in the CR and the apparent inversion of the PSM cannot be overlooked. We have concluded that these methodological ambiguities reiterate the call for public access to the underlying Korean National Health Insurance database for independent validation and resolution of any contradiction.

4.2. The Epidemiological Paradox and Lack of External Validity

The core issue addressed by this paper has been the epidemiological paradox which demonstrates a profound lack of external validity for the Kim HJ et al. cohort. This is evidenced by the suppressed overall CR (40.78 per 10,000) compared to the national average (52.46 per 10,000), which indicates that the final sample is strongly suspected to be not representative of the underlying population's cancer incidence.

This fundamental inconsistency also suggests that the sampling bias present in the initial cohort selection was not adequately corrected by the Propensity Score Matching, or that the matching procedure itself introduced a new, significant bias. The lack of representativeness of the Kim HJ et al. cohort, evidenced by the suppressed overall CR, calls into question the reliability of the study's conclusions. The discrepancy is so large (11.68 per 10,000) that it cannot be dismissed as a minor statistical (or procedural) fluctuation.

4.3. PSM Inversion as a Plausible Cause of Bias

The hypothesis that the Propensity Score Matching (1:4 PSM) was inverted provides a concrete methodological explanation for the observed low CR and the resulting lack of external validity. The standard procedure is to use the Treatment Group (Vaccinated) as the reference for building the Control Group (Unvaccinated). Given the reported final 1:4 ratio skewed towards the vaccinated group, the matching appears to have used the smaller unvaccinated pool as the base, thus inverting the procedure. This inversion could potentially lead to severe methodological consequences, specifically suggesting two major downstream issues that could have compromised the results' integrity. They are the following. First, a compromised control group potentially leads to the selection of an unrepresentative sub-sample of the smaller, unvaccinated population pool, jeopardazing the integrity of the control group against which all comparisons are made. Second, the result is an exacerbated sampling bias which manifests as an inversion that may either introduce unidentified confounding factors or fail to balance known ones effectively. This, in turn, would result in an artificial suppression of the overall cancer incidence rate in the final cohort compared to the national reality.

Nonetheless, we must clarify that our intention is not to reject the statistical associations found by Kim HJ et al. in [3], showing a higher incidence of cancers in the COVID-19 vaccinated population. Such associations, while potentially valid *within the limits of their non-representative cohort*, are not the focus of this critique. However, it is impossible to ignore the profound discrepancies in the Crude Incidence Rate and the strong evidence suggesting the inverted application of the 1:4 PSM. These possible methodological flaws would render the external validity and, consequently, the reliability of the hazard ratios derived from this specific cohort, highly questionable until proven otherwise.

4.4. Comparison with National Epidemiological Standards

The fundamental premise of any large population-based study is that the raw incidence rate of any common background event (like cancer) must be consistent with national epidemiological surveillance for the same time period. The 22.26% downward deviation observed in the Kim HJ et al. cohort's CR constitutes a plausible evidence of its lack of external validity. This failure is particularly alarming given the study's reliance on a Propensity Score Matching procedure, which, theoretically, should *improve* external validity, not degrade its consistency with the national baseline. The fact that the overall CR is so low despite the use of PSM suggests that either the initial cohort selection could be flawed and cannot be rescued by statistical matching, or the matching itself was performed incorrectly, as we hypothesize.

4.5. Limitations

The limitations inherent to this critique stem primarily from the nature of observational analysis built upon aggregated, externally published data. Specifically, our quantitative findings regarding the suppressed Crude Incidence Rate (CR) and the hypothesis of Propensity Score Matching (PSM) inversion cannot be definitively resolved without primary data access. However, these limitations are ultimately attributable to the original Kim et al. study, which, despite its vast scale, did not provide the necessary methodological transparency or data access for independent validation. The ethical and scientific imperative for data transparency remains the single greatest constraint, forcing critical analysis to rely on numerical signatures and logical inference rather than direct verification.

5. Conclusions

The Kim HJ et al. study [3] presents a significant result that, due to possible methodological and numerical inconsistencies, cannot be accepted without deeper scientific scrutiny. The paradoxical low overall cancer incidence rate and the strong evidence suggesting an inverted PSM procedure could undermine the integrity of the cohort, unless proven otherwise. To resolve these ambiguities and reinforce scientific rigor, the following actions are essential: a) transparency from authors who should disclose the detailed algorithmic methodology and group assignment used for the PSM, and b) full disclosure of the initial data to allow for independent verification and resolution of the raised methodological ambiguities. While public availability of the entire Korean National Health Insurance database is ideal, a comprehensive, extended summary of the raw data, consistent with privacy regulations, would be acceptable, provided it is detailed enough to confirm or invalidate the methodological concerns raised here. Resolving the identified methodological ambiguity is fundamental to the global trust in reported associations and broader COVID-19 vaccine safety assessments.

Supplementary Materials: No further supporting information.

Author Contributions: Not applicable. M.R. as a single author has contributed to all stages of the production of this manuscript and has read and agreed to its published.

Funding: This research received no external funding.

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

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

Data Availability Statement: All data used for calculations are publicly available and explicitly cited from the Kim HJ et al. study's supplementary materials and official Korean cancer registries. All calculations are easily reproducible.

Acknowledgments: The author is grateful to several colleagues who provided comments on previous preprints on this matter.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CR	Crude incidence Rate
ASR	Age Standardized Rate
N	Number
PSM	Propensity Score Matching
SD	Standard Deviation
WHO	World Health Organization

References

1. Chirico F, Teixeira da Silva JA. (2022). COVID-19 Health Policies: The need for transparent data sharing etween Scientists, Governments, and Policymakers. *Oman Med J*, 37(5). doi: 10.5001/omj.2022.63.
2. Moro PL, Haber P, McNeil MM. (2019). Challenges in evaluating post-licensure vaccine safety: observations from the Centers for Disease Control and Prevention. *Expert Rev Vaccines,* 18(10):1091–1101. doi: 10.1080/14760584.2019.1676154.
3. 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.
4. The Strait Times Editor, South Korea opens Covid-19 vaccine reservations for all adults, (2023) (August). *The Straits Times*.

https://www.straitstimes.com/asia/east-asia/south-korea-opens-covid-19-vaccine-reservations-for-all-adults (accessed online 15 October 2025*)*.

5. Paul E, Steptoe A, Fancourt D, (2021). Attitudes towards vaccines and intention to vaccinate against COVID-19: Implications for public health communications. *The Lancet Regional Health - Europe*, 1: 100012. doi: 10.1016/j.lanepe.2020.10001.

6. Roccetti M. (2025). A Critical Note on Contradictions in South Korean Cancer Incidence Rates: The Paradox of Crude Rates Derived from the Kim HJ et al. Cohort (Biomark Res, 13:114, 2025) Showing Concurrent Increases in the Vaccinated and Overall Decrease. *Preprints.org*, 2025. doi: 10.20944/preprints202510.0883.v1.

7. Roccetti M. (2025). Addendum to "A Critical Note on Contradictions in South Korean Cancer Incidence Rates: The Paradox of Crude Rates Derived from the Kim HJ et al. Cohort (Biomark Res, 13:114, 2025)" - On Possible Sampling Bias and Inverted Propensity Score Matching. *Preprints.org*, 2025. doi: 10.20944/preprints202510.1664.v1.

8. 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.

9. 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.

10. 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.

11. Last JM (2014). *A Dictionary of Public Health*. Oxford University Press. doi: 10.1093/acref/9780195160901.001.0001.

12. WHO, (2023). Republic of Korea: Health data overview. *World Health Organization*, https://data.who.int/countries/410 (accessed online 15 October 2025).