Human Interaction and Emerging Technologies (IHIET-FS 2026), Vol. 235, 2026, 149–156
https://doi.org/10.54941/ahfe1007270

Structural Risk Recalibration and Stochastic Stationarity in Localized Large-Scale Health Data: An Intelligent Difference-in-Differences Framework

Marco Roccetti

University of Bologna, Bologna, I-40126, Italy

ABSTRACT

This study proposes a framework for analyzing the reliability of spatial health data through structural system integrity. We define a health system state through a tuple *S* = <*L*, *C*, *D*, *E*, *NE*, *OBS*, *RF*, *RRF*>, where local observations (*L*) must be calibrated against a national contextual baseline (*C*) to ensure inferential integrity. In this architecture, a portion of the experimental population (*E*) is exposed to a specific determinant (*D*), compared to a non-exposed (*NE*) group. By applying four mathematical operators, a total system gap (*Delta*), a relative weighting gap (*G_w*), a structural difference-in-differences function (*DiD*), and a recentered influence function (*RIF*), we executed our context-aware framework on a large-scale health-disease incidence local dataset relative to its broader spatial context. Our framework demonstrated its efficacy by extrapolating, in the example of interest, a critical 32.5% *G_w*, indicating an under-representation of the older population within the local space compared to the national context. Significantly, the application of *DiD* uncovered a 18.2% asymmetric divergence of the observed disease incidence in the *NE* (non-exposed) group compared to *E*, relative to national benchmarks. The subsequent *RIF* recalibration failed to reach a contextual leverage, proving that the initially calculated risk factor (*RF*) was an outcome of internal fractures rather than a signal emitted by the health system. The use of our framework proved that an unbalanced choice of the experimental population (marked by a –45.1% deficit in the older *NE* arm) had rendered the initial calculation of risk structurally unstable. We demonstrate that without contextual calibration, geospatial large scale health data can produce phantom signals indistinguishable from systemic errors.

Keywords: Structural health remodelling, Recentered influence function, Health risk recalibration, Spatial difference-in-differences, Inferential stability

INTRODUCTION

Local health risk signals, characterized by high-magnitude associations and alarming statistical significance, frequently originate from limited observational data streams. When these signals are fueled by massive N-size datasets, the dimensional scale of the sample may induce an interpretation of local evidence as a prominent natural hazard warning. This study demonstrates that such risk configurations often constitute stochastic distortions generated by the spatial information of the original local repositories that, despite their

Received March 15, 2026; Revised March 30, 2026; Accepted April 17, 2026; Available online June 22, 2026

size, are unanchored from more general surrounding realities (Roccetti et al., 2019). We present a mathematical framework and prove its efficacy by resolving a specific phenomenon of local health risk signal inflation through a structural remodelling procedure of risk associations. By integrating a recentered influence function (*RIF*) regression, our model expands the analytical context beyond the boundaries of the local dataset. The process systematically examines historical metrics and national-level standards, allowing for a rigorous reconsideration of the observed risk parameters. Through this application of our intelligent spatial difference-in-differences (*DiD*) decomposition, the local signal is partitioned into an explained component, justified by the distribution of local covariates, and a structural unexplained component. This decomposition isolates the divergence between the observed phenomena within the dataset and the national baseline, highlighting how the magnitude of the local health alarm is a function of a structural incidence deficit internal to the local data architecture. Our modelling demonstrates that the original alarm may not derive from actual independent variables, but from the local repository-specific stochastic configuration. A further employment of inferential statistics verifies the divergence of local data properties from national-context stationarity. Such results confirm that risk associations triggering critical alarms within closed local computational cycles can be systematically mitigated and returned to non-alarming values once recalibrated against extra-local scenarios (Kruskal et al., 1979; Van Delden et al., 2018; Casini et al., 2021).

In the particular scrutinized case, we operationalized our framework through the analysis of a specific geospatial health state, stepping through three different phases (Oaxaca, 1973; Blinder, 1973; Prosperi et al., 2020; Marfia and Roccetti, 2010). First, we formalized the health system state by defining the local observation mechanism (L) relative to the national contextual standard (C). Second, we used a *DiD* operator to detect whether the failure in the local data was unevenly distributed between that portion of the experimental population exposed (E) to a given determinant (i.e, COVID-19 vaccination) and the non-exposed (NE) portion. Third, and final, we tried to recalibrate the local risk factor (RF) of the observed disease (i.e., cancer) by projecting it onto the national demographic and incidence stationarity.

Summarizing, our framework on the specific used dataset has revealed that the initial reported alarm was the outcome of a 32.5% structural weight deficit in the elderly stratum of the experimental population along with an 18.2% asymmetric divergence of the observed disease incidence in the NE (non-exposed) group compared to E and relative to national benchmarks. When the *RIF* repair operator was applied, the local signal (33.43 cancer cases per 10,000) failed to align with the national ground truth (55.02 cases per 10,000). Our framework proved that the NE portion of the experimental population was the result of an unbalanced choice in the extraction of that local sub-group (NE) as marked by a –45.1% deficit in the number of cases of occurrence of the diseases compared to the national standard. In essence, in the investigated case, the initially calculated risk factor (RF) was an outcome of internal fractures rather than a pure health signal extensible to

a broader context. Our research establishes that RIF-based recalibration can be an indispensable procedural requirement for the validation of any large-scale yet local risk assessment, ensuring that realistic warnings are rooted in the reality of historical and national data. By isolating the structural difference between the local observational frequency and the extra-local incidence mean, we effectively neutralize the architectural distortions of the local repository.

We conclude by noting that a phenomenon of large-N localized bias is not unique in the recent literature. Similar structural fractures were already observed in a UK biobank case (Fry et al., 2017), where the massive scale of the dataset (N = 500,000) initially produced alarming risk associations that were later mitigated when compared to the general UK population's socio-economic and mortality stationarity. Similarly, during 2021, high-velocity local health data suggested rapid waning of clinical determinants in a case in Israel (Goldberg et al., 2021). Subsequent structural audits revealed that the *NE* groups were fundamentally decoupled from the national baseline due to differential healthcare-seeking behavior, creating an elusive risk signal similar to the one identified in our dataset. The remainder of the paper proceeds as follows. In the next Section, we provide a definition for a spatial health system and we develop our theoretical framework to analyze it. In the subsequent Section, we provide the results we achieved by applying our framework to a very exemplar case, while the final Section terminates our paper with a brief discussion.

GEOSPATIAL HEALTH SYSTEMS: A THEORETICAL COMPUTATIONAL FRAMEWORK

We consider a geospatial health system (S) situated in a specific local domain (L) and nested within a broader national context (C). The system is observed through a dataset of large dimensions N and is characterized by the following state tuple:

$$S = < L, C, D, OBS, OBS\_E, OBS\_NE, RF, RRF >.$$

Where: D is the determinant, that is the external factor introduced into the health system (exposure), for example a massive COVID-19 vaccination extended all over the local experimental population. OBS are the observations of the occurrence of a given disease (e.g., cancer) following D. Those detected effects (cases), are structured in a dual-level hierarchy. Goups: Exposed (E) and Non-Exposed (NE) and Strata: Young ($young < 65$) and Elderly ($older >= 65$). RF is the computed local risk factor, that is the internal metric of the local system's response, while RRF is the recalibrated risk factor, in simpler terms the corrected state after recalibrating local observations with the contextual reference C. Notable is the fact that in this formalization, any discrepancy between L and C is not to be interpreted as a biological issue, but as the permanence in the system of a non-integrity state, the final objective being of analyzing whether the local observation mechanism accurately reflects the contextual reality. As

anticipated, the validity of a health system state can be determined by its structural integrity, assessed through four mathematical operators: **1**) the total system gap (*Delta*), measuring the absolute distance between the local state and the contextual baseline, that is *Delta* = *RF_obs* - *RF_exp*, which is the difference between the risk factors calculated either based on local observations or on the baseline proposed by the broader national context. This operator detects possible initial biases. If *Delta* does not equal 0, the system *L* is operating in a decoupled state from context *C*, suggesting the local observation mechanism is fundamentally inconsistent with the national baseline. Then we have: **2**) the relative weighting gap (*G_w*), measuring the sampling distortion of the disease-high-incidence stratum (*older*), with the *w* parameters indicating the portions of elderly in the local experimental population compared to the national baseline: *G_w* = 1 – ((*w_*(*L*, *older*) / *w_*(*C*, *older*)). This formula exposes the compositional error. In health systems where effects are concentrated in a specific age bracket of a population, a high *G_w* proves the local system lacks the structural mass required to produce a valid inference compared to the national reference. Now it arrives **3**) the structural difference-in-differences function (*DiD*). This starts from the measurement of the relative deviation of the local component (*L*) from the context (*C*) by comparing the observed disease incidences in the local health system against the national context: *Diff_L* = (*OBS_L* - *OBS_C*) / *OBS_C*. At this point, the *DiD* is not merely a statistical result, but the direct algebraic identity emerging from the subtraction of the observational efficiencies of the two arms, considering the portions of population exposed or not exposed to the determinants (vaccination): *DiD* = *Diff_*(*L*, *E*) – *Diff_*(*L*, *NE*). In essence, any non-zero *DiD* value would prove that the observation is asymmetric, that is that one arm of the local dataset (the *NE*) can be more broken or less monitored than the other, creating a mechanical bias in the risk calculation. This is a very relevant check. If it fails (*DiD* not zero) this means that the initial risk factor is not a property of the determinant *D*, but a byproduct of this mathematical asymmetry within those chosen portions of the experimental population. When a health system fails to maintain an equal balance across all groups, a false difference is guaranteed to appear. The risk is not measuring a health effect; it is measuring the spatial decay of its own observation integrity. Finally **4**), that is the recentered influence function (*RIF*). It is the repair operator, attempting to reconstruct a *RRF* by projecting local observations onto the contextual distribution, with the formula: $RRF = \sum OBS\_L \times w\_C$. The resulting value represents the mathematical convergence between local observation (*OBS_L*) and national stability (*w_C*). This product acts as a filter designed to test the robustness of an anomaly. It ensures that a local signal is considered valid only if the injection of national weight produces a reinforced product capable of reaching the system's ground truth. If the recalibrated product fails to reach the national benchmark, despite this systemic amplification, the local incidence is immediately exposed as a geospatial artifact. An insufficient product zeroes out the initial risk, proving that the national support is inadequate to validate the local peak. Failing to bridge the gap *L* / *C* identifies the risk as a loss of systemic integrity.

EXPERIMENTAL RESULTS

We instantiate the *L*, *C* contexts and the *D* determinant, with all their relative values, in terms of local and national disease incidences, and original data from the following references: (Kang et al., 2023; Kim et al., 2025; Park et al., 2024; Park et al., 2025; Statista, 2024; World Bank, 2024), which are summarized in Table 1.

Table 1: Consolidated data summary.

Group	*L – OBS* (Incidence/10k)	*C – EXP* (Incidence/10k)	Deficit (*Diff* -)	Weight (*w_L*)	Weight (*w_C*)
>= 65 y, *NE*	85.2	155.2	45.1%	12.15%	18%
>= 65 y, *E*	113.5	155.2	26.9%	12.15%	18%
H. System	33.43	55.02	39.24%	100%	100%

We now execute our analysis to determine if the local *RF* (initially computed equal to 1.21, after PSM-derived adjustments) deviates from the national context using the four operators mentioned earlier.

We begin by calculating *Delta* with our formula for the difference between the local and national risk ratios: *Delta* = 1.21 - 0.7748 = 0.4352. Note that 0.7748 was achieved dividing the disease incidence of *E* at *L* 42.63 (Kim et al., 2025) by the expected incidence of *NE* 55.02 based on *C*. This value is a measure of the decoupled state (local vs. national). It shows that starting from a local alarm of a 21% risk increase for the *E*, relying solely on the isolated context of the N-size dataset, we arrive at almost 0.77, when benchmarked against the *C*, with an expected risk deficit of almost 23% and a structural discrepancy of over 43 percentage points between local perception and systemic reality.

As a subsequent step, we applied the four structural operators exposed in the previous Section. First, we run the *G_w*, measuring the mass of the system. The local system *L* contains only 12.15% elderly subjects, while the context *C* requires 18.00%, hence: *G_w* = 1 – (0.1215/0.18) = 0.325 (32.5%). This explains that the investigated health system is structurally compromised by a 32.5% volume deficit in its most critical observation stratum (older population). In essence, this results indicates that in the original N size dataset nearly one-third of the population that should have driven the results (elderly) were silenced.

Third, we applied the *DiD* function testing for asymmetry in the detection mechanism of the elderly stratum, contrasting *E* and *NE*, yielding: *DiD* = (–26.9% – (–45.1%)) = 18.2%. This 18.2% gap proves that the *NE* arm of the experimental population was significantly more broken than the *E* arm. The initial risk factor is thus likely to be not due to the determinant *D* (vaccination) but to this 18.2% asymmetry. In essence, the *NE* arm of the experimental population was monitored so poorly (missing 45.1% of cancer cases) compared to *E* (missing 26.9%) that a difference was mathematically guaranteed to appear well beyond the role of the determinant.

Finally, we attempted to repair the system by reweighting the *NE* observations to match the 18.00% weight of the national *C*, using our *RRF* operator and considering that: for the entire (younger + older) experimental population *OBS_*(*L*, *NE*) = 31.2 (per 10,000), *OBS_C* = 55.02 (per 10,000) and *w_*(*C*, *NE*) = 18%. Hence, *RRF* = (31.2 × (1 +18.2)) = 36.88. In synthesis, the recalibrated value remained after recalibration decoupled from the contextual ground truth (36.88 vs. 55.02). This was the final evidence. Even after mathematically injecting the missing elderly individuals to reach the national 18% weight (*RRF* = 36.88), the incidence was still nowhere near the real national level of 55.02. This means the problem was not just the quantity of people, but that the selected subjects were artificially healthy. Consolidating the gap between 36.88 and 55.02 in the absolute measure of the bias confirms the distance between the local source of data *L*, regardless of its large dimensions, and the broader national context *C*.

CONCLUSION WITH DISCUSSION

We begin by eliminating any ambiguity regarding the origin of the employed parameters and data presenting an extended tabulation of all the data, reported in Table 2.

Table 2: Master data resources.

Category	National *C*	Local *NE*	Local *E*
Elderly (w >= 65)	18.00%	12.15%	12.15%
Young (w < 65)	82.00%	87.85%	87.85%
Participants (Young)	N/A	522,722	2,090,888
Participants (Elderly)	N/A	72,285	289,140
Total Participants	N/A	595,007	2,380,028
Cases (Young)	N/A	1,373	6,861
Cases (Elderly)	N/A	616	3,283
Total Cases	N/A	1,989	10,144
Incidence Rate (Young)	33.03 / 10k	26.27 / 10k	32.81 / 10k
Incidence Rate (Elderly)	155.20 / 10k	85.22 / 10k	113.54 / 10k
Total Incidence Rate	55.02 / 10k	33.43 / 10k	42.63 / 10k

Our structural context-aware analysis led to three results. First, we revealed that the 32.5% *G_w* created an initial decoupling of *L* from the national context. By under-representing the high-risk elderly stratum by nearly a third, the local baseline is fundamentally shifted. While the reporting compliance is locally present, inferential integrity, that is the ability of the local dataset to support the conclusion is near zero. Importantly, the most critical finding is the 18.2% *DiD* gap. In the local elderly stratum, the observation mechanism of the *NE* is deflated by 45.1% relative to the national *C*, while the *E* group is subject to a negative variation by only 26.9%. More crucially, the *RIF* proves that the health system is beyond repair. Even when we correct the age weights to match the national 18%, the resulting incidence (36.88) remains 33% lower than the national ground truth (55.02).

In the end, the initial alarming risk (cancer increase following vaccination) is a mathematical byproduct of the *NE*'s inability to capture cases, being unjustified to compare an unreliable control group with a less unreliable exposed group claiming that the difference is biological. To conclude, we have shown, running an examplar case, that a structural analysis based on difference-to-difference techniques may be useful to assess the reliability of geospatial health systems identifying the differences between what emerges from the local data, regardless of its relative size and a broader context.

We would like also to emphasize that the case analyzed here has served only as a single exemplary instance, and further cases should be tested to definitively validate the effectiveness of the proposed framework (Ferretti et al., 2007; Marfia et al., 2011; Marfia et al., 2010, Palazzi et al., 2010; Roccetti, 2026). Furthermore, we reiterate our complete agnosticism regarding the vaccination-cancer debate, whether affirmative or negative, which remains within the scope of the examined data but outside the role of our analysis.

REFERENCES

Blinder, AS. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *J. Hum. Resour.* 8(4):436–455. DOI:10.2307/144855.

Casini, L. Marchetti, N. Montanucci, A. et al. (2023) A human–AI Collaboration Workflow for Archaeological Sites Detection. *Sci. Rep*. 13:8699. DOI: 10.1038/s41598-023-36015-5.

Ferretti, S. Mirri, S. Roccetti, M. et al. (2007) Notes for a Collaboration: On the Design of a Wiki-type Educational Video Lecture Annotation System. *Proceedings of the International Conference on Semantic Computing*, 2007. 651–656. DOI: 10.1109/ICSC.2007.18.

Fry, A. Littlejohns, TJ. Sudlow, C. et al. (2017) Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol. 186*(9):863–888. DOI: 10.1093/aje/kwx246.

Goldberg, Y. Mandel, M. Bar-On, YM. et al. (2021) Waning Immunity after the BNT162b2 Vaccine in Israel. *N. Engl. J. Med*. 385(24):e85. DOI: 10.1056/NEJMoa2114228.

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.

Kim, HJ. Kim, M-H. Choi, MG. et al. (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.

Kruskal, V. Mosteller, F. (1979) Representative sampling, I: Non-scientific literature. *Int. Stat. Rev.* 47(1):13–24. DOI: 10.2307/1403202.

Marfia, G. Roccetti, M. (2010) TCP at Last: Reconsidering TCP's Role for Wireless Entertainment Centers at Home. IEEE Trans. Consum. Electron. 56(4):2233–2240. DOI: 10.1109/TCE.2010.5681095.

Marfia, G. Roccetti, M. Amoroso, A. et al. (2011) Cognitive cars: Constructing a Cognitive Playground for VANET Research Testbeds. ACM International Conference Proceeding Series. 2011:29. DOI: 10.1145/2093256.2093285.

Oaxaca, R. (1973) Male-Female Wage Differentials in Urban Labor Markets. *Int. Econ. Rev*. 14(3):693–709. DOI:10.2307/2525981x.

Palazzi, CE. Roccetti, M. Marfia, G. (2010) Realizing the Unexploited Potential of Games on Serious Challenges. *Comput. Entertain.* 8(4):23. DOI: 10.1145/1921141.1921143.

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.

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

Prosperi M, Guo Y, Sperrin M, et al. (2020). Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare. *Nat. Mach. Intell.* 2:369–375. DOI: 10.1038/s42256-020-0197-y.

Roccetti, M. (2026) Before the algorithm: An Exemplar Case of the Necessity of Statistical Testing for Epidemiological Consistency in Public Health Data. *AIMS Public Health.* 13(1):121–134. DOI: 10.3934/publichealth.2026008

Roccetti, M. Delnevo, G. Casini, L. et al. 2019 Is Bigger always Better? A Controversial Journey to the Center of Machine Learning Design, with Uses and Misuses of Big Data for Predicting Water Meter Failures. *J. Big Data.* 6(1):70. DOI: 10.1186/s40537-019-0235-y.

Statista. South Korea: Cancer Crude Incidence Rate by Age (2024) [Accessed 2026 Mar 10]. Available from: https://www.statista.com/statistics/1440818/south-korea-cancer-crude-incidence-rate-by-age/.

Van Delden, A. van Der Laan, J. Prins, A. 2018 Detecting Reporting Errors in Data from Decentralised Autonomous Administrations with an Application to Hospital Data. *J. Off. Stat.* 34(4):863–888. DOI: 10.2478/jos-2018-0043.

World Bank. Population ages 65 and above (% of total population) - Korea, Rep. World Population Prospects, United Nations (UN), (2024) [Accessed 2026 Mar 10]. Available from: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?locations=KR.