Human Systems Engineering and Design (IHSED 2026), Vol. 237, 2026, 673–680
https://doi.org/10.54941/ahfe1008206

Human-Centric Health Data Fidelity: Inferential Methodologies for Calibrating Human Systems Integration in Global Health Surveillance

Marco Roccetti

University of Bologna, Bologna, I-40126, Italy

ABSTRACT

We start from the idea that the reliability of global health surveillance on health data in the present era depends on the expert's ability to validate the fidelity of massive data through rigorous inferential procedures. This paper discusses the potential of a formal framework for assessing the fidelity of a Health Data Space (S), where a determinant (D) precedes monitored effects (F) within a large-scale dataset of exposed (E) and non-exposed (NE) individuals of a given experimental population. We propose a methodology to determine whether risk alerts represent global biological phenomena or unstable local signals by anchoring the data of a health space to established national gold standards and well known references. Specifically, our methodology employs a four-way inferential sequence based on Chi-squared Goodness-of-Fit tests to detect structural deviations across four critical dimensions: demographic proportional representation, clinical fidelity of high-risk strata, internal consistency of the unexposed baseline, and systemic divergence of aggregate incidences of the health space. We demonstrate the efficacy of this framework through a real-world case study of a massive dataset, specifically selected for its known structural complexities and its impact on the international public health discourse. The results we achieved confirm that this strategy is capable of identifying terminal divergences, such as demographic skews and biologically impossible deficits, proving that reported signals can often be products of a data selection bias rather than a biological effect. This approach confirms the necessity of human-centric validation, providing a formal mathematical protocol to calibrate global health systems and ensure that surveillance remains grounded in demographic and empirical reality.

Keywords: Human-centric data fidelity surveillance, Health data space, Signal stability, Inferential calibration

INTRODUCTION

The digital transformation of global health has led to the emergence of massive, interconnected health data spaces, where the representation of human pathology is no longer confined to clinical records but is ingested into automated surveillance loops. In this evolving landscape, the reliability of such systems does not merely depend on the volume of data, the so-called Big Data paradigm, unprecedented in its scale, but on the humans' ability

Received March 30, 2026 Revised June 13, 2026 Accepted July 7, 2026 Available online September 25, 2026

to validate the fidelity of the digital signals through rigorous mathematical control techniques, like inferential statistics above all (Roccetti et al., 2019). This validation is technically grounded in the necessity of maintaining an anchor to real-world reality, moving beyond descriptive summaries to a formal testing of the hypotheses extracted by those massive health datasets. Within this framework, the human expert is asked to perform several non-delegable functions that are essential for the safe and ethical operation of health surveillance systems. Clinical validation, for instance, allows clinicians to oversee algorithmic outputs, intervening at crucial moments to confirm diagnoses and prevent the propagation of false positives or negatives in ambiguous clinical cases. Furthermore, bias mitigation requires direct human intervention to analyze whether the system is performing equitably across different subpopulations defined by age, ethnicity, or gender, preventing disparities in care. This contextual decision-making enables the integration of machine-generated insights with the patient's individual clinical history and broader reference values (Casini et al., 2021; Obermeier et al., 2019; Topol, 2019). However, this structural integration can be fundamentally compromised when the data architecture is not inferentially calibrated, leading to a scenario where the system operates in a vacuum, detached from the biological and demographic constraints of the population it intends to represent (Ghassemi et al., 2020; Kelly et al., 2019; Marfia and Roccetti, 2010).

To evaluate the fidelity of such systems, we discuss a formal Health Data Space (S) as a multidimensional domain where we identify a Determinant (D), representing a specific public health intervention or exposure, which chronologically precedes, though does not necessarily cause, a set of monitored Effects (F). The stability of the resulting signals is analyzed by partitioning the population into two primary strata: the Exposed (E), consisting of individuals within the vast sample who have encountered the determinant D, and the Non-Exposed (NE), which serves as the control baseline of individuals not subjected to the intervention. The core of our methodology lies in determining whether the observed effects F are manifestations of global systemic (i.e., biological) phenomena or merely unstable local signals. In a human-centric health system, the relationship between E and NE must remain anchored to the demographic and clinical history of a broader population, the history of which is composed of established age distributions, disease incidences, and recognized clinical gold standards. The human expert's role in this framework is to ensure that the data space does not collapse into a distorted vision of reality, if the transition from D to F occurs within a fractured sampling of E and NE. In fact, if the NE baseline represents a biologically impossible outlier or if E is demographically skewed, the resulting signal could lose its inferential validity. In ths context, with our research we propose a four-way strategic framework based on χ^2 Goodness-of-Fit tests to detect structural deviations across four critical dimensions: demographic proportional representation, clinical fidelity of high-risk strata, internal consistency of the NE baseline, and systemic divergence of aggregate incidences of the health space. By benchmarking system outputs against national population parameters and

well-known references, we provide the human expert with the necessary tools to detect structural sampling biases before they are formalized into misleading public health signals (Alba et al., 2017; Prosperi et al., 2020; Roccetti, 2026). Ensuring that global health surveillance remains a trustworthy collaboration between human expertise and machine processing is fundamental to the future of complex systems interaction and global health governance. To demonstrate the efficacy of this framework, we have applied our methodology to a real-world case study involving a massive dataset specifically selected for its known structural complexities and its impact on international public health discourse. The results confirm that this 4-way strategy has been capable of identifying terminal divergences, proving that reported signals can often be products of data selection bias rather than biological reality, and preventing the exportation of unstable findings. The remainder of the paper proceeds as follows. In the next Section, we provide our methodological framework. In the subsequent Section, we offer the results we achieved by applying our framework to a very exemplar case, while the final Section terminates our paper.

HEALTH DATA FIDELITY: INFERENTIAL METHODOLOGIES FOR GLOBAL SURVEILLANCE

The evaluation of fidelity within a Health Data Space (S) requires a rigorous transition from descriptive observation to inferential validation. In this framework, we define a Determinant (D) as the primary exposure which was specifically identified in an experimental context (for example a COVID-19 vaccination campaign,) which precedes the monitoring of Effects (F) (here represented by the incidence of oncological diagnoses). The population is partitioned into the Exposed (E) group, consisting of vaccinated individuals, and the Non-Exposed (NE), serving as the control baseline. The test of the stability of these signals was anchored to a relevant real world case, for which all the relative data (even though in an aggregated form) was made available, along with all the national references extracted from high-fidelity national registries. Technically, our methodology was based on the Pearson's χ^2 Goodness-of-Fit test. The χ^2 has been chosen as the primary inferential tool because it measures the discrepancy between observed frequencies (O) in local health data spaces and the expected frequencies (Ex) derived from broader (e.g., national) population benchmarks. From a mathematical viewpoint, the standard formula is $\chi^2 = \sum\left(O(i) - Ex(i)\right)^2 / Ex(i)$. Where χ^2 is the statistical value representing the total divergence; O(i) are the observed frequencies (the actual data from the health space) and Ex(i) are the expected frequencies (the value that should exist if the data from the local health space would perfectly match national standards). In this context, a high χ^2 value signifies that the dataset defining a given health data space is no longer a representative sample but has undergone a structural deviation from the entire population. Our methodology utilizes this type of test across a four-way strategic analysis, where each phase acts as a possible sensor of system instability (Pencina and D'Agostino, 2015). The first phase (**Test 1**) aims at measuring the demographic proportionality, that is it evaluates whether the

age distribution of the health data space, specifically the ratio between the elderly (≥ 65 years) and younger strata, remains congruent with the national census. A deviation here woud indicate a selection bias that would inherently distort any age-dependent disease incidence. The second phase (**Test 2**) aims at testing the data fidelity of high-risk strata; in essence it scrutinizes the health status of the most vulnerable demographic (the elderly) to ensure their recorded incidence of F is consistent with expected clinical expectations. The third phase (**Test 3**) wishes to test the internal consistency of the unexposed baseline, the NE group, to verify that the control portion extracted from S is not a healthy-survivor phenomenon, which would artificially inflate the perceived risk in the opposite E group. Finally, with **Test 4** we develop a systemic convergence analysis, by performing a triple-layer validation of the aggregate health space. In essence, we compare not only the total health space's incidence against the national average value of incidence of F (i.e., Crude Incidence Rrate or CR), but it also tests the Exposed (E) and Non-Exposed (NE) groups individually against this same incidence standard. This granular approach is designed to detect if the determinant D is operating within a fractured environment where both the treatment and control arms have drifted towards statistically impossible trajectores. As already anticipated, we have employed our methodology in a real world high-complexity health setting situated in the Asia-Pacific region, with the *S* space and the *D* determinant (COVID-19 vaccination) and all their relative figures in terms of local and national disease incidences (cases of cancer) extracted from a massive dataset whose original data can be retrieved using the following references: (Kang et al., 2023; Kim et al., 2025; Park et al., 2024; Park et al., 2025; Statista, 2024; World Bank, 2024). Those most relevant data are summarized in Table 1.

Table 1: Structural integrity matrix: demographic and incidence divergences.

Phase	Parameter	Nation	Space S	Difference	% Gap
Test 1	≥ 65 %	18.0%	12.15%	–5.85%	–32.5%
	< 65 %	82.0%	87.85%	+5.85%	+7.1%
Test 2	CR Age ≥ 65 (per 10k)	155.20	107.9	–47.30	–30.5%
Test 3	CR NE ≥ 65 (per 10 k)	155.20	85.20	–70.00	–45.1%
Test 4	Aggregate Total - CR (per 10,000)	52.46	40.78	–11.68	–22.3%
	Exposed (E) - CR (per 10,000)	52.46	42.63	–9.83	–18.7%
	Non-Exposed (NE) – CR (per 10,000)	52.46	33.43	–19.03	–36.3%

With respect to the data portrayed in Table 1, it is worth adding only that, given the biological nature of oncogenesis (F) and the latency required for clinical detection, the official national Crude Rate (CR) chosen for the comparison was averaged over years 2020-2022 showing a value of 52.46 per 10,000 (SD 2.97).

RESULTS: INFERENTIAL TESTING OF THE HEALTH DATA SPACE

The transition from a descriptive representation of health data space S to an inferential validation requires a methodology capable of identifying structural inconsistencies between that S and its population of reference. Our four-way strategic analysis on the scrutinized case has acted as a high-precision sensor, extracting all the systemic inconsistencies across demographic and clinical dimensions of the original dataset. By instantiating the χ^2 formula for each critical phase, we achieved the results shown in Table 2. The accuracy of our methodology lies in the ability to anchor the local health space to official population benchmarks, ensuring that no digital abstraction can mask a systemic deviation. In every instance, our analysis has successfully isolated a distinct structural issue.

Table 2: Inferential test matrix: χ^2 calculation and outcomes.

Group	χ^2 Formula	χ^2 Result	Outcome
Test 1	$(361{,}425 - 535{,}506)^2 / 535{,}506$	77,654.2	Structural demographic deviation
Test 2	$(3{,}899 – 5{,}609)^2 / 5{,}609$	728.6	Data selection divergence
Test 3	$(85.20 – 155.20)^2 / 155.20$	914.7	Data integrity instability
Test 4 (Total)	$(40.78 – 52.46)^2 / 52.46 \times N$	7,748.1	Divergence S/Nation (Total)
Test 4, E	$(42.63 – 52.46)^2 / 52.46 \times$ size(E)	4,961.4	Divergence S/Nation (E)
Test 4, NE	$(33.43 – 52.46)^2 / 52.46 \times$ size(NE)	2,835.6	Divergence S/Nation (NE)

In particular, with Test 1 (demographic proportionality), our procedure demonstrated its accuracy by converting the –5.85% absolute difference (a –32.5% relative gap) into a massive population deficit. By applying both the proportion of 12.5% from space S and the national benchmark of 18% to the entire dimension of the health space (comprised of N = 2,975,035 individuals), the analysis extracted a missing population of 174,039 elderly individuals (361,425 observed vs. 535,506 expected). This specific extraction revealed that the health space S had effectively de-aged the surrounding reality. The resulting extreme χ^2 (> 77,000) proves that this is not a sampling fluke but a structural exclusion that mechanically deflates F incidences by thinning out the primary age-related drivers. With Test 2 (Incidence fidelity) applied to our exemplar case, our methodology has then revealed that demographics were not the only skewed variable. Focusing on the elderly individuals present

in S, the audit extracted a –30.5% deficit in cancer incidence (107.9 vs. 155.2). By exploiting the combination of the demographic rates and the CR frequencies, the procedure identified a deficit of 1,710 clinical cases (3,899 observed vs. 5,609 expected). This outcome has proved that the original data space included atypically healthy individuals, diverging from the biological morbidity recorded by national registries. The χ^2 of 728.6 has exposed a Healthy Patient Effect so profound that the elderly in S must be considered biologically different from the general population (Chemaitelly et al., 2025). With Test 3 (Internal space integrity), instead, our methodology focused on the NE elderly arm. By identifying a deficit of 70 points of incidence (–45.1%) compared to the national baseline of 155.20, the procedure has proved that the NE represented a void of cases. The χ^2 of 914.7 confirmed that this baseline was too healthy w.r.t. standards. Finally, the Test 4 (Systemic convergence) performed a triple-layer extraction of the aggregate drift of the entire health space S against the national standard CR reference of 52.46. Specifically, by comparing the average space F incidence (40.78) against the national CR, the analysis identified a –22.3% gap. The resulting χ^2 (7,748.1) exposed a total space drift proving that the initial massive dataset was not a mirror of the national reality. E, alone, instead, has shown an incidence of 42.63, with a –18.7% deviation from the same benchmark. The χ^2 value (4,961.4) in this case proved that even the active arm of S was deflated, likely due to the demographic thinning identified in Test 1. Finally, the NE. The most critical deviation is emphasized here (χ^2 = 2,835.6). The NE in fact has shown an incidence of only 33.43, a terminal –36.3% gap from the national standard. This result was the definitive evidence of an unstable data health environment: the NE arm is so far removed from systemic (biological) reality that it creates a vacuum, making the E appear at high risk only because the baseline was statistically impossible. It is also worth noting that all the tests were performed under a p-value < 0.001 condition, while size (E) and size (NE) can be easily calculated with the information of Table 1, also remembering the value of N = 2,975,035. In conclusion, our exercise has shown that our 4-way methodology identifies relevant deviations even in a complex data system architecture like that investigated in this study. In particular, Test 4 confirmed that S was not a representative sample of a broader reality but a stochastic oscillation where the E/NE arms have drifted into impossible territories. We would like to reiterate that the analyzed case served only as a single exemplary instance, while other cases should be analyzed to confirm the efficacy of the methodology (Ferretti et al., 2007; Marfia et al., 2011; Marfia et al., 2010, Palazzi et al., 2010). Finally, we maintain our complete agnosticism regarding the vaccination-cancer debate, whether affirmative or negative.

CONCLUSION

The methodology presented in this paper has been structured around a defined health data space, characterized by a specific determinant and a biological effect, articulated across two distinct arms of the experimental population. To validate its integrity, we developed a multi-layered global health sureveillance methodology based on the χ^2 statistic, to detect systematic discrepancies

between the local health space and broader national references. This diagnostic procedure was engineered to monitor four critical dimensions: i) Demographic proportionality, ii) Data fidelity, iii) Baseline symmetry, and iv) Systemic convergence. Applied to a highly complex data architecture, the procedure demonstrated exceptional analytical performance. By extracting exact localized divergences with extreme statistical precision.

REFERENCES

Alba, AC. Agoritsas, T. Walsh, M. et al. (2017). Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA*. 318(14):1377–84. DOI: 10.1001/jama.2017.12126.

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.

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.

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.

Ghassemi, A. Naumann, T. Schulam, P. et al. (2020) A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Jt. Summits Transl. Sci. Proc*. 2020:191–200. PMID: 32477638.

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):385–99. DOI: 10.4143/crt.2023.447.

Kelly, JK. Karthikesalingam, A. Suleyman, M. et al. (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17(1):195. DOI: 10.1186/s12916-019-1426-2.

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.

Marfia, G. Roccetti, M. (2010) TCP at Last: Reconsidering TCP's Role for Wireless Entertainment Centers at Home. *IEEE Trans. Consum. Electron*. 56(4):2233–40. 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.

Obermeyer, R. Powers, B. Vogeli, C. et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 366(6464):447–53. DOI: 10.1126/science.aax234.

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):357–71. 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):312–30. DOI: 10.4143/crt.2025.264.

Pencina, JM. D'Agostino, JB. Sr. (2015). Evaluating Discrimination of Risk Prediction Models: The C Statistic. *JAMA*. 314(10):1063–4. DOI: 10.1001/jama.2015.11082.

Prosperi, M. Guo, Y. Sperrin, M. et al. (2020). Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare. *Nat. Mach. Intell.* 2:369–75. DOI: 10.1038/s42256-020-0197-y.

Roccetti, M. (2026) Quantifying structural selection bias in observational cohort data: A ponderation analysis of age - specific incidence rates to inform vaccine safety verification. *Front. Pharmacol.* 16:1754809. DOI: 10.3389/fphar.2025.1754809.

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/.

Topol, M. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25:44–56. DOI: 10.1038/s41591-018-0300-7.

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.