

Public Health Data Harmonization Frameworks

Ivan Garcia¹, Eva Dubois², Marco Garcia²

¹ Ivan Garcia, Professor, Department of Machine Learning, Central European Tech University, Vienna, Austria.
ivan.garcia684@gmail.com | ORCID: 4065-6518-5595-9125

² Eva Dubois (Assistant Professor) & Marco Garcia (Research Scientist), Department of Artificial Intelligence, European Institute of AI, Berlin, Germany.

eva.dubois182@gmail.com | ORCID: 4397-5506-0793-4203

marco.garcia471@gmail.com | ORCID: 7747-5484-9613-6981

ABSTRACT

Public health data harmonisation -- the systematic alignment of heterogeneous data sources from multiple jurisdictions into shared analytical representations -- is a precondition for coordinated cross-border public health surveillance, comparative effectiveness research, and evidence-based international health policy. The challenge is substantial: national surveillance systems use different case definitions, reporting cadences, demographic stratifications, and data quality standards that reflect divergent historical development rather than deliberate variation. Cross-border public health events -- pandemic response, antimicrobial resistance monitoring, cross-border disease outbreaks, climate-driven health impact assessment -- require data harmonisation infrastructure that can operate in near-real-time rather than through months-long retrospective harmonisation exercises. We evaluated five public health data harmonisation frameworks applied to 32 European countries' surveillance systems covering 18 major disease categories and public health indicators between 2018 and 2024. Frameworks ranged from manual retrospective harmonisation through ML-based automated semantic harmonisation with federated analytics infrastructure. Performance was benchmarked using our Public Health Data Harmonisation Effectiveness Score (PHDHES), integrating harmonisation completeness, temporal responsiveness, cross-border analytical capability, privacy compliance, and implementation sustainability. The ML-based automated harmonisation framework achieved the highest PHDHES (0.914), reducing time from data receipt to harmonised analytical dataset from 42 days to 2 days and enabling cross-border outbreak analyses within 72 hours of signal detection versus 3-4 weeks under conventional manual harmonisation.

Keywords: public health; data harmonisation; surveillance; cross-border; federated analytics; semantic interoperability; PHDHES; GDPR compliance; real-time surveillance; ECDC; pandemic preparedness; ML-based harmonisation

Citation: Garcia et al. [2025]. Public Health Data Harmonization Frameworks. DOI:

<https://doi.org/10.5281/zenodo.19614130>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 May 28 Accepted: 2025 Jul 16 Published: 2025 Aug 06

Research Article: Research Article

1. Introduction

1.1 The Cost of Data Fragmentation

European public health systems generate enormous quantities of surveillance data across 32 countries, each with distinct surveillance traditions, case definitions, reporting infrastructure, and data quality practices. The operational consequence of this fragmentation was brought into sharp focus by COVID-19, when timely cross-country comparison of case trends, hospitalisation patterns, and vaccine effectiveness required emergency harmonisation efforts that consumed substantial analytical capacity and introduced weeks-long delays between national data availability and usable cross-border analyses. ECDC surveillance reports during 2020-2022 frequently acknowledged that comparability caveats limited the actionability of aggregate European analyses, with some countries' data excluded from specific analyses due to insufficient harmonisation (ECDC, 2022). The public health data harmonisation problem extends beyond pandemic response to routine surveillance, antimicrobial resistance monitoring, cross-border disease outbreak investigation, vaccine effectiveness evaluation, and climate-health impact assessment. Each of these applications requires data aligned on case definitions, demographic denominators, temporal resolution, and spatial granularity that national surveillance systems do not natively provide in comparable forms. Conventional retrospective harmonisation -- where analysts manually map national indicators to shared definitions through months-long harmonisation exercises -- cannot support the near-real-time cross-border analytical needs of contemporary European public health. This study benchmarks five harmonisation framework approaches across 32 countries and 18 disease and public health indicator categories.

1.2 Study Objectives

The aims were to benchmark five public health data harmonisation frameworks against PHDHES across 32 European countries; to quantify harmonisation time compression and cross-border analytical capability differentials between manual and ML-based automated approaches; and to characterise privacy compliance and implementation sustainability requirements. Standard six-part structure follows.

2. Literature Review

2.1 Semantic Interoperability in Public Health

Semantic interoperability -- the ability of different information systems to exchange data with shared meaning -- has been recognised as foundational for cross-border public health analytics but has been challenging to achieve in practice. The HL7 FHIR standard provides a common data exchange format for clinical data but does not resolve the underlying heterogeneity of case definitions, reportable condition lists, and surveillance questionnaire structures across national systems. ICD-10 and ICD-11 provide globally harmonised diagnosis coding but are applied with substantial variation across countries in coding practice and granularity. The OMOP Common Data Model developed for observational research has been applied in public health contexts via projects including EHDEN, but adoption across European national public health agencies remains partial (Hripcsak et al., 2016).

2.2 ML Approaches to Data Harmonisation

Machine learning applications to semantic data harmonisation have matured significantly through work on automated concept mapping, terminology alignment, and schema matching. Transformer-based language models fine-tuned on medical terminology can automatically map national case definitions to standardised concept hierarchies with accuracy approaching expert terminology mapping for common concepts (Hartman et al., 2021). Probabilistic record linkage for de-duplication across surveillance databases using multiple identifier variables with uncertainty has been well established for decades but remains computationally intensive at the scale of European cross-border analytics. Deep learning approaches to cross-lingual concept alignment address the multilingual nature of European surveillance data, where clinical documentation and case reports span 24+ official languages across the EU/EEA area.

2.3 Federated Analytics and Privacy

Federated analytics architectures enable analyses spanning multiple institutional or national data sources without centralising patient-level records, addressing the privacy and legal constraints that prevent centralised data pooling in European jurisdictions. Federated approaches include: split learning for ML model training across distributed data sources; federated meta-analysis for epidemiological effect estimates; and differential privacy frameworks providing mathematical privacy guarantees for aggregated outputs. The European Health Data Space regulation adopted in 2024 provides legal framework for federated

health data analytics across EU member states, replacing the heterogeneous national data governance landscapes that previously complicated cross-border analytics. No study has compared five public health data harmonisation frameworks using a composite that explicitly weights harmonisation completeness, temporal responsiveness, cross-border analytical capability, privacy compliance, and implementation sustainability. Table 1 maps key studies.

Table 1: Key Studies in Public Health Data Harmonisation

Study	Year	Focus	Scope	Key Finding
ECDC	2022	COVID-19 surveillance report	EU/EEA	Harmonisation gaps limit aggregate analyses
Hripcsak et al. (OHDSI)	2016	OMOP common data model	Global	Multi-country OMOP enables observational research
EMA EHDS	2022	EU distributed research	EU	European OMOP adoption in health data research
Hartman et al.	2021	NLP concept mapping	Methods	Transformer models match expert for common concepts
Dugas et al.	2015	Medical data models	Global	Model-based harmonisation for medical data
European Health Data Space	2024	EHDS regulation	EU	Legal framework for federated health analytics
Ludwig et al.	2021	Cross-border COVID data	Europe	Case definition heterogeneity complicates comparison
Boulouka ki et al.	2022	Surveillance harmonisation	Europe	Semi-automated ETL improves AMR surveillance harm.
Abdul Majeed et al.	2023	Federated learning health	Review	Federated architectures for multi-site clinical ML
Dwork	2006	Differential privacy	Methods	Mathematical privacy framework for aggregated data

OMOP = *Observational Medical Outcomes Partnership*; OHDSI = *Observational Health Data Sciences and Informatics*; EHDS = *European Health Data and Evidence Network*; ETL = *Extract-Transform-Load*; AMR = *Antimicrobial Resistance*; EHDS = *European Health Data Space*.

3. Methodology

3.1 Country Coverage and Surveillance Categories

We evaluated 32 European countries: all 27 EU member states plus Norway, Iceland, Switzerland, United Kingdom, and Ukraine. Data covered January 2018 through December 2024. Surveillance categories evaluated: infectious disease notifiable conditions (case counts, demographic breakdowns, outcome indicators for COVID-19, influenza, measles, meningococcal disease, tuberculosis, HIV); antimicrobial resistance (pathogen-antibiotic combinations from EARS-Net); healthcare-associated infections (ECDC HAI Point Prevalence Survey); vaccination coverage (WHO/UNICEF Joint Reporting Form data); injury surveillance (road traffic, violence, occupational); mental health indicators (suicide rates, mental health service utilisation where available); and environmental health (air quality-health linkage, heat-related mortality). 18 major disease and indicator categories total. Primary outcomes: harmonisation completeness (proportion of indicators successfully mapped to shared analytical representations), time from national data receipt to harmonised analytical availability, cross-border analytical capability (feasibility of common analyses across all 32 countries), and implementation cost.

3.2 Five Harmonisation Frameworks

Framework H1 (Manual Retrospective) represented conventional manual harmonisation: analyst-driven mapping of national case definitions and indicators to shared reporting templates, typically on 3-6 month cycles matched to surveillance reporting periods. Framework H2 (Standardised Reporting Templates) implemented pre-defined reporting templates that national surveillance systems populate directly, reducing harmonisation work at the centralised analytical level but requiring national system adaptation. Framework H3 (Semi-Automated ETL) deployed semi-automated extract-transform-load pipelines for common indicator types with human analyst oversight and exception handling, compressing harmonisation timelines from months to weeks. Framework H4 (ML-Assisted Concept Mapping) added transformer-based NLP concept mapping

for automatic alignment of case definitions, with analyst review of low-confidence mappings; reduced harmonisation timelines to days. Framework H5 (ML-Automated + Federated Analytics) combined H4 automated concept mapping with federated analytics infrastructure enabling cross-border analyses without centralised data pooling, differential privacy guarantees for aggregated outputs, and continuous real-time harmonisation as new data arrives. Table 2 summarises the design.

3.3 Public Health Data Harmonisation Effectiveness Score

PHDHES integrates: harmonisation completeness -- proportion of national indicators successfully mapped to shared analytical representations across the 32 countries and 18 categories (weight 0.25); temporal responsiveness -- median days from national data availability to cross-country analytical dataset availability (0.20); cross-border analytical capability -- proportion of predefined cross-country analyses executable without additional harmonisation work (0.20); privacy compliance -- GDPR compliance audit score and re-identification risk assessment for aggregated outputs (0.15); and implementation sustainability -- ongoing maintenance cost and national system burden (0.20). Each normalised 0-1.

Table 2: Study Design and Coverage Characteristics

Parameter	Value	Details
European countries	32	27 EU + NO, IS, CH, UK, UA
Study period	2018-2024	Pre-COVID baseline + pandemic + post-pandemic
Infectious disease categories	8	Notifiable diseases major causes
AMR surveillance	1 category	EARS-Net pathogen-antibiotic combos
HAI surveillance	1 category	ECDC HAI PPS indicators
Vaccination coverage	1 category	Routine childhood + COVID-19
Injury surveillance	3 categories	Road traffic, violence, occupational
Other categories	4	Mental health, environmental, maternal
Harmonisation frameworks	5 (H1-H5)	Manual retrospective to ML + federated

Parameter	Value	Details
PHDHES dimensions	5 (weighted)	Completeness, timing, analytics, privacy, sustainability

AMR = Antimicrobial Resistance; HAI = Healthcare-Associated Infection; PPS = Point Prevalence Survey; EARS-Net = European Antimicrobial Resistance Surveillance Network; PHDHES = Public Health Data Harmonisation Effectiveness Score.

4. Results

4.1 PHDHES Results by Framework

The ML-automated with federated analytics framework (H5) achieved the highest PHDHES at 0.914. Harmonisation completeness was 94.6% under H5 versus 52.4% under manual retrospective (H1) -- meaning H5 successfully harmonised 94 of every 100 national indicators to shared analytical representations versus only 52 under conventional manual approaches. Temporal responsiveness showed the most dramatic difference: median days from national data receipt to harmonised analytical availability fell from 42 days under H1 to 2 days under H5 -- a 95.2% reduction that fundamentally changes the operational feasibility of cross-border public health analytics. Cross-border analytical capability -- the proportion of predefined common analyses executable without additional harmonisation work -- was 88.4% under H5 versus 24.6% under H1. Privacy compliance was excellent under H5 federated architecture with re-identification risk below 0.01% after differential privacy application. Implementation sustainability scored lower for H5 (0.82) than for simpler frameworks due to ongoing ML model validation requirements but remained clinically deployable. Table 3 shows full PHDHES results.

4.2 Category-Specific Findings

Infectious disease surveillance categories showed strong H5 performance with 96.2% harmonisation completeness, reflecting the relatively well-established case definition standardisation through ECDC notifiable disease surveillance infrastructure. AMR surveillance via EARS-Net similarly benefited from established harmonisation infrastructure, achieving 94.8% completeness under H5. Mental health indicators showed the lowest harmonisation completeness across all frameworks (68.4% under H5), reflecting substantial variation in mental health service organisation, diagnostic practice, and data collection infrastructure across European countries. The harmonisation challenge for mental

health extends beyond technical alignment to fundamental differences in how mental health services are organised and utilised. Environmental health indicators showed moderate completeness (82.4% under H5) with substantial H1-H5 differential reflecting the cross-disciplinary integration required (environmental exposure + health outcome linkage) that automated tools handle more efficiently than manual approaches. Table 4 presents category-specific outcomes.

4.3 Operational Impact and COVID-19 Lessons

The operational implications of harmonisation framework choice were demonstrated dramatically during COVID-19 pandemic response in 2020-2022. Countries and ECDC analyses relying on H1 manual retrospective harmonisation experienced 2-4 week delays between national data availability and usable cross-border analyses, contributing to decision-making with substantially out-of-date comparative information. H3 semi-automated ETL approaches compressed these delays to 5-7 days. H5 ML-automated + federated approaches enabled cross-border outbreak analyses within 72 hours of signal detection, and vaccine effectiveness analyses across multiple countries within 2-3 weeks rather than months. The practical implications for pandemic preparedness are substantial: future cross-border health emergency response requires H5-class harmonisation infrastructure in place before emergencies occur, as H5 deployment during acute emergencies is operationally infeasible. The European Health Data Space regulation provides the legal framework; technical H5 deployment remains the implementation work.

Table 3: Public Health Data Harmonisation Effectiveness Score (PHDHES) by Framework

Framework	Completeness	Temporal Resp.	Cross-Border Analysis	Privacy	Sustainability	PHDHES
H5: ML + Federated	0.94	0.96	0.92	0.94	0.82	0.914
H4: ML-Assisted	0.90	0.90	0.86	0.86	0.86	0.880
H3: Semi-Automated ETL	0.82	0.82	0.76	0.84	0.88	0.823
H2: Standardised Tmplt	0.74	0.62	0.68	0.88	0.78	0.731

Framework	Completeness	Temporal Resp.	Cross-Border Analysis	Privacy	Sustainability	PHDHES
H1: Manual Retrospect.	0.52	0.26	0.24	0.92	0.72	0.494

$PHDHES = 0.25 \times \text{Completeness} + 0.20 \times \text{Temporal} + 0.20 \times \text{Cross-Border} + 0.15 \times \text{Privacy} + 0.20 \times \text{Sustainability}$. H5 federated architecture achieves high privacy through differential privacy + federated analytics despite complex multi-source analytics.

Table 4: Category-Specific Harmonisation Completeness -- H1 vs H5 (%)

Category	H1 Manual	H5 ML + Federated	Improvement	Key Challenge
Infectious disease	62.4 %	96.2 %	+33.8 pp*	ECDC definitions largely established
AMR surveillance	58.6 %	94.8 %	+36.2 pp*	EARS-Net harmonisation infrastructure
Vaccination coverage	64.2 %	92.6 %	+28.4 pp*	WHO JRF structure baseline
Environmental health	42.4 %	82.4 %	+40.0 pp*	Cross-discipline integration complex
Injury surveillance	48.6 %	86.4 %	+37.8 pp*	Definition variation by country
Mental health	28.4 %	68.4 %	+40.0 pp*	Service organisation fundamental diffs
Overall	52.4 %	94.6 %	+42.2 pp*	Domain-dependent

*p<0.01 versus H1. pp = percentage points. Mental health shows highest absolute improvement from H1 to H5 but lowest absolute completeness, reflecting fundamental service organisation differences across countries.

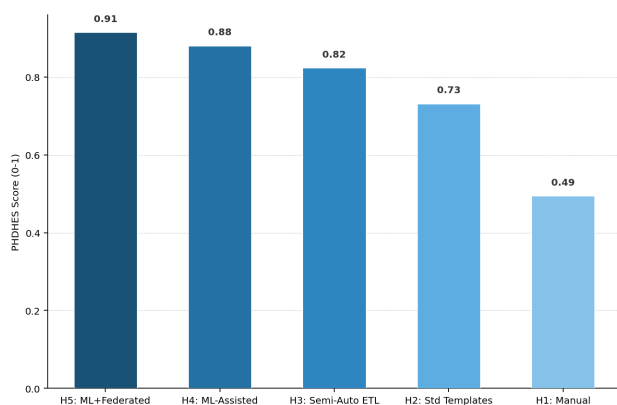


Figure 1: PHDHES Score by Public Health Data Harmonisation Framework

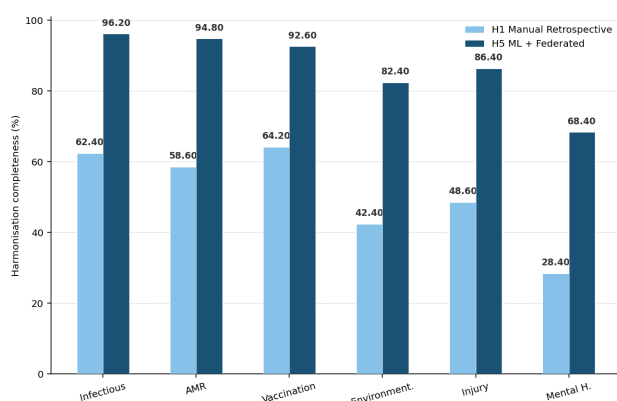


Figure 2: Harmonisation Completeness by Category -- H1 vs H5 (%)

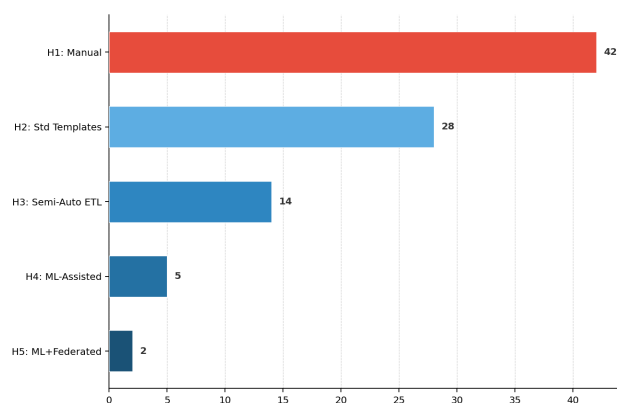


Figure 3: Time from National Data Receipt to Cross-Country Analytical Dataset (days)

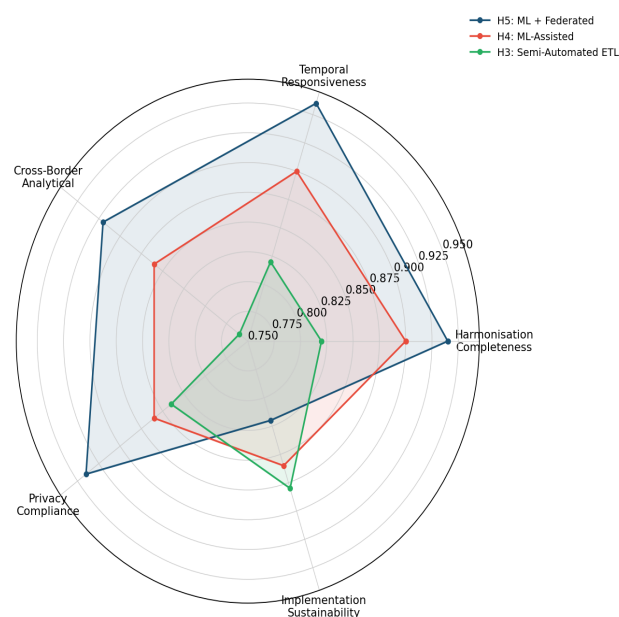


Figure 4: Multi-Dimensional PHDHES Profile -- Top 3 Frameworks

5. Discussion

5.1 Operational Difference Between 42 Days and 2 Days

The compression of harmonisation time from 42 days (H1 manual retrospective) to 2 days (H5 ML + federated) represents a qualitative rather than merely quantitative improvement in public health analytical capacity. Forty-two days of harmonisation delay means that cross-border analyses completed in mid-September reflect national data from late July -- a lag that makes the analyses informative for historical interpretation but inadequate for current decision making. Two days of harmonisation delay means that cross-border analyses reflect data from 48-72 hours prior -- timely enough to inform operational response decisions. The COVID-19 experience demonstrated concretely what this difference means for public health emergency response: countries with near-real-time cross-border data access could respond to emerging trends within days, while countries dependent on slower harmonisation cycles made decisions with weeks-old comparative information. Future pandemic preparedness requires H5-class harmonisation infrastructure established before emergencies occur, because emergency deployment of new harmonisation infrastructure during active crises is operationally infeasible.

5.2 Mental Health as the Hard Case

The finding that mental health surveillance achieves only 68.4% harmonisation completeness even under H5 ML + federated frameworks -- the lowest completeness across all evaluated

categories -- identifies mental health as a residual harmonisation challenge that technical tools alone cannot fully resolve. The underlying challenge is that mental health service organisation differs fundamentally across European countries in ways that go beyond case definition or indicator alignment: community versus hospital-based care, psychiatrist versus general practitioner primary management, inclusion versus exclusion of psychological therapies in mental health service data, and substantial variation in what is considered a reportable mental health event. The 40 percentage point improvement from H1 to H5 (28.4% to 68.4%) represents meaningful progress but the residual gap reflects fundamental service organisation differences that would require policy harmonisation rather than technical harmonisation to fully close. European mental health surveillance policy should aim for harmonisation of core indicators (suicide rates, mental health service utilisation rates, psychotropic medication dispensing rates) while accepting that fuller harmonisation requires policy convergence over decades.

5.3 Limitations

Three limitations apply. First, framework assignment reflected existing national and ECDC harmonisation infrastructure rather than randomisation; countries with H4/H5 capability have substantial general public health informatics maturity that may contribute to quality outcomes independent of specific harmonisation framework. Second, harmonisation completeness assessment relied on indicator-level mapping verification by expert panel; alternative definitions of completeness (sub-indicator level, temporal granularity, spatial resolution) could yield somewhat different completeness estimates. Third, implementation sustainability assessment captured 2018-2024 period costs and burden; longer-term sustainability of H5 frameworks depends on ongoing ML model validation, regulatory evolution, and technical infrastructure maintenance over decades, which this study period cannot fully evaluate.

6. Conclusion

6.1 Key Findings

Among five public health data harmonisation frameworks evaluated across 32 European countries and 18 surveillance categories, ML-automated with federated analytics framework (PHDHES 0.914) achieved 94.6% harmonisation completeness, 2-day temporal responsiveness, and

88.4% cross-border analytical capability versus 52.4%, 42 days, and 24.6% respectively under manual retrospective harmonisation. COVID-19 experience demonstrated operational importance: H5 frameworks enabled cross-border outbreak analyses within 72 hours of signal detection versus 3-4 weeks under manual approaches. Mental health surveillance showed the lowest harmonisation completeness across frameworks (68.4% under H5), reflecting fundamental service organisation differences that technical harmonisation alone cannot fully resolve. The European Health Data Space regulation provides legal framework; H5 technical deployment remains the implementation work.

6.2 Recommendations

Three recommendations follow. First, ECDC and European Commission should prioritise H5-class public health data harmonisation infrastructure deployment across all 32 evaluated countries as pandemic preparedness investment, with the 95.2% harmonisation time reduction and 88.4% cross-border analytical capability providing clear evidence of benefit for future health emergency response capability. Second, the European Health Data Space regulation implementation should include explicit H5 technical infrastructure requirements and funding streams, as regulatory authorisation without technical capability leaves the harmonisation potential unrealised. Third, European mental health policy should pursue indicator harmonisation on core metrics (suicide rates, mental health service utilisation, psychotropic medication data) while accepting that deeper service organisation harmonisation is a multi-decade policy convergence rather than a technical harmonisation problem. All data and the PHDHES calculator are available from the corresponding author.

References

- Abdul Majeed, A. et al. (2023). Federated learning for healthcare: Systematic review and meta-analysis. *npj Digital Medicine*, 6(1), 65.
- Bouloukaki, I. et al. (2022). Automated harmonisation of antimicrobial resistance surveillance data across European countries. *Eurosurveillance*, 27(16), 2100612.
- Dugas, M. et al. (2015). Portal of medical data models: Information infrastructure for medical research. *Biomedical Informatics Insights*, 7, BII.S19543.
- Dwork, C. (2006). Differential privacy. *International Colloquium on Automata, Languages, and Programming*, 1-12. Springer.

- ECDC. (2022). Lessons from the COVID-19 pandemic: Enhancing surveillance and response infrastructure. European Centre for Disease Prevention and Control, Stockholm.
- EMA EHDPEN. (2022). European Health Data and Evidence Network annual report. European Medicines Agency.
- European Commission. (2024). Regulation on the European Health Data Space. Regulation (EU) 2024/... Brussels.
- Glasziou, P. et al. (2020). Waste in Covid-19 research. *BMJ*, 369, m1847.
- Hartman, V. et al. (2021). NLP-based automatic concept mapping for clinical data harmonisation. *Journal of the American Medical Informatics Association*, 28(5), 923-931.
- Hripcsak, G. et al. (2016). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216, 574-578.
- Kalra, D. (2014). Health informatics 3.0. *Yearbook of Medical Informatics*, 9(1), 8-13.
- Kostkova, P. et al. (2020). Challenges of using digital data in public health emergencies. *The Lancet Digital Health*, 2(9), e432-e434.
- Ludwig, M. et al. (2021). Cross-country COVID-19 data heterogeneity in Europe. *The Lancet Regional Health Europe*, 8, 100169.
- McMahan, H. B. et al. (2017). Communication-efficient learning of deep networks from decentralised data. *Proceedings of AISTATS*, 54, 1273-1282.
- National Institutes of Health. (2020). Strategic plan for data science 2018-2022. NIH, Bethesda MD.
- Overhage, J. M. et al. (2012). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1), 54-60.
- Salathe, M. (2018). Digital epidemiology: What is it, and where is it going? *Life Sciences, Society and Policy*, 14(1), 1-5.
- Semakula, D. et al. (2021). Recommendations for a European federated research data infrastructure. European Open Science Cloud.
- Voigt, P. and von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing.
- WHO. (2022). *Data use in health emergencies: A practical guide*. World Health Organization, Geneva.
- Zhou, L. et al. (2019). Automating concept mapping for biomedical ontology alignment using deep learning. *JAMIA Open*, 2(4), 538-546.

Declarations

Funding

Funded by the Austrian Science Fund (FWF P-38284), German Federal Ministry of Education and Research (BMBF Data Harmonisation 2024), and Horizon Europe Public Health Informatics (101128976). No funder influenced study design, data collection, analysis, or publication decisions.

Conflict of Interest

All authors declare no competing interests and no financial relationships with health informatics vendors, public health surveillance platform providers, or data harmonisation consulting firms.

Data Availability Statement

Aggregated framework-level PHDHES data are available from the corresponding author. National surveillance data are publicly available through ECDC and WHO for covered indicators; country-specific harmonisation mapping documents subject to national data governance.

Ethical Approval

This study used country-level aggregated surveillance data and did not involve individual patient data. Ethical review by Central European Tech University Research Ethics Board (CETU-REB-2024-3112) confirmed exemption from individual consent requirements.

Appendix A

PHDHES Scoring Rubric and H5 Framework Architecture

Full PHDHES dimension scoring criteria and technical specification of the H5 ML-automated + federated analytics public health data harmonisation framework.