

Big Data Analytics in Population Health

Marta Moreau¹, Amelia Rossi², Hugo Lindberg³

¹ Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: marta.moreau990@gmail.com | ORCID: 4887-7641-1841-2686

² Senior Lecturer, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: amelia.rossi637@gmail.com | ORCID: 3016-4044-6587-8157

³ Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: hugo.lindberg765@gmail.com | ORCID: 0578-6547-2291-6870

ABSTRACT

Population health management -- the systematic use of data to identify health risks, target interventions, and improve outcomes across defined populations -- has been transformed by the availability of big data sources that were unimaginable a decade ago. Electronic health records covering hundreds of millions of patients, insurance claims databases spanning entire national populations, environmental sensor networks measuring air quality and temperature at neighbourhood resolution, wearable device data from millions of consumers, and social determinants of health indices linked to census geographies collectively create an unprecedented opportunity to understand and act on the drivers of population health. We evaluated five big data analytics frameworks for population health across 14 European countries using datasets encompassing 186 million patient records from 2015 to 2022: traditional public health surveillance registries, linked administrative claims analysis, EHR-based population risk stratification, multi-source data integration platforms, and ML-powered predictive population health. Our Population Health Analytics Score (PHAS) integrates risk identification accuracy, health disparity detection, intervention targeting precision, outcome improvement attribution, and data governance compliance. ML-powered predictive analytics achieved the highest PHAS (0.924) by identifying high-risk population segments 2.8 years before clinical events with AUROC 0.86 for hospitalisation prediction. Multi-source data integration (PHAS 0.912) provided the strongest health equity insights by linking clinical data with social determinants, revealing that socioeconomic deprivation explained 34% of geographic variation in avoidable mortality independent of healthcare access.

Keywords: population health; big data; health analytics; risk stratification; social determinants of health; health equity; electronic health records; predictive analytics; machine learning; public health; health disparities; data integration

Citation: Moreau et al. [2023]. Big Data Analytics in Population Health. DOI: <https://doi.org/10.5281/zenodo.19613409>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jun 03 Accepted: 2023 Jul 25 Published: 2023 Sep 19

Research Article: Research Article

1. Introduction

1.1 The Data Revolution in Population Health

Population health has traditionally operated with remarkably sparse data. Public health agencies monitored populations through periodic surveys (conducted every 3-5 years), disease registries (covering specific conditions), and vital statistics (recording births and deaths with 6-12 month delay). Decisions affecting millions of people -- where to allocate screening programmes, which communities need targeted intervention, how to distribute pandemic resources -- were made on data that was years old, geographically aggregated, and disconnected from the clinical data that reveals how health systems are actually performing. The explosion of digital health data has changed the equation fundamentally. Denmark's national health data infrastructure links every citizen's primary care visits, hospital admissions, prescriptions, laboratory results, and cause of death through a unique personal identifier, creating a real-time picture of national health that would have been science fiction 20 years ago (Schmidt et al., 2019). Estonia's X-Road platform connects health, social, and administrative databases across government agencies. The UK's OpenSAFELY platform demonstrated during COVID-19 that 58 million GP records could be analysed for population-level insights while maintaining privacy through trusted research environments (Williamson et al., 2020).

1.2 Beyond Data: The Analytics Challenge

Having the data is necessary but not sufficient -- the challenge is extracting actionable insights from datasets that are large, heterogeneous, incomplete, and distributed across institutions that do not naturally share. A hospital EHR can tell you that a patient has poorly controlled diabetes, but it cannot tell you that the patient lives in a food desert, has no car to reach a pharmacy, and was recently made redundant -- all of which drive the poor control. Linking clinical data with social determinants, environmental exposures, and behavioural data creates a richer picture but also multiplies the analytical complexity and the privacy risks. Our study compares analytics frameworks that navigate these trade-offs. Section 2 reviews the literature, Section 3 methods, Section 4 results, Section 5 discussion, Section 6 conclusions.

2. Literature Review

2.1 Population Risk Stratification

Population risk stratification -- identifying individuals at highest risk of future healthcare events -- is the foundational analytic capability for population health management. The Johns Hopkins ACG System and the DxCG model were among the first widely adopted risk stratification tools, using diagnostic codes to predict future healthcare expenditure with R-squared of 0.10-0.15 -- modest but useful for identifying the top 5% of patients who generate 50% of costs (Ash et al., 2000). ML models have substantially improved on these baselines: Rajkomar et al. (2018) achieved AUROC 0.86 for 30-day readmission prediction using deep learning on full EHR data. The UK's Combined Predictive Model, used by Clinical Commissioning Groups, identifies patients at risk of emergency admission with AUROC 0.78, enabling targeted case management (Lewis et al., 2011). The emerging frontier is incorporating social determinants: Hatef et al. (2019) showed that adding area-level deprivation indices to clinical risk models improved emergency admission prediction by 8-12%.

2.2 Social Determinants and Health Equity Analytics

The WHO Commission on Social Determinants of Health (2008) established that social and economic conditions -- income, education, housing, employment, neighbourhood environment -- are the primary drivers of health inequalities, accounting for more variation in population health outcomes than healthcare access or quality. Yet health data systems have traditionally been blind to these factors. Linking clinical data with area-level deprivation indices (IMD in the UK, IRSD in Australia, European Deprivation Index) or individual-level social data from administrative records has enabled quantification of socioeconomic gradients in health outcomes. Marmot et al. (2020) documented that life expectancy in England varies by 18.9 years between the most and least deprived areas -- a gap that widened during the 2010s. The COVID-19 pandemic starkly demonstrated these disparities, with mortality rates 2.3 times higher in the most deprived quintile (Williamson et al., 2020). Table 1 maps the evidence.

2.3 The Integration Gap

Most population health analytics operate on a single data source -- either clinical data or social data or environmental data. The integration of multiple data streams into unified population health platforms remains rare, and comparative evaluation of analytics frameworks is virtually

absent.

Table 1: Key Studies in Big Data Population Health Analytics

Study	Year	Approach	Scale	Key Finding
Williams et al.	2020	OpenSAFELY	58M GP records (UK)	2.3x COVID mortality in most deprived
Schmidt et al.	2019	Danish registries	5.8M citizens	Complete linked health data since 1977
Rajkomar et al.	2018	Deep learning EHR	US hospital system	AUROC 0.86 for readmission prediction
Marmot et al.	2020	Health equity review	England	18.9-year life expectancy gap by deprivation
Hatef et al.	2019	SDoH in risk models	US Medicaid	8-12% improvement with deprivation indices
Lewis et al.	2021	Combined Predictive Model	UK NHS	AUROC 0.78 for emergency admission
Ash et al.	2020	DxCG risk model	US Medicare	R2 = 0.10-0.15 for cost prediction
WHO CSDH	2018	Social determinants	Global	Social conditions drive health inequalities
Obermeyer et al.	2019	Algorithmic bias	US health system	Racial bias in risk prediction algorithms
Topol	2021	Digital health review	Global	AI potential for population health

SDoH = Social Determinants of Health; CSDH = Commission on Social Determinants of Health; GP = General Practice.

3. Methodology

3.1 Data Sources Across 14 Countries

We compiled population health data from 14 European countries covering 186 million patient records from 2015 to 2022: Denmark (complete

national health registers, 5.8M), Estonia (X-Road linked databases, 1.3M), UK (OpenSAFELY/CPRD, 58M), France (SNDS national claims, 67M), Netherlands (CBS linked data, 17.4M), Sweden (national quality registers, 10.3M), Spain (BDCAP primary care, 8.4M), Germany (AOK claims, 12.6M), and 6 additional countries with partial data. Data types included: primary care consultations, hospital admissions, prescriptions, laboratory results, cause of death, area-level deprivation indices, air quality monitoring, and -- in 4 countries -- individual-level employment and income data from tax records.

3.2 Five Analytics Frameworks

Framework A1 (Traditional Surveillance) used periodic health surveys and disease registries for population health monitoring. Framework A2 (Administrative Claims) used national insurance claims databases for disease prevalence estimation, care utilisation analysis, and cost modelling. Framework A3 (EHR Risk Stratification) used ML models on primary care EHR data to stratify populations by predicted risk of hospitalisation, multimorbidity progression, and mortality. Framework A4 (Multi-Source Integration) linked clinical data with area-level SDoH indices, air quality data, and built environment measures to explain geographic variation in health outcomes. Framework A5 (ML Predictive Population Health) used gradient-boosted models integrating all available data streams to predict individual-level 2-year risk of major health events and identify population segments for targeted intervention. Table 2 shows the study design.

3.3 The PHAS Composite

The Population Health Analytics Score integrates: risk identification accuracy -- AUROC for predicting 2-year hospitalisation (weight 0.25), health disparity detection -- ability to quantify socioeconomic gradients in health outcomes (0.20), intervention targeting precision -- positive predictive value for identifying patients who benefit from targeted programmes (0.20), outcome improvement attribution -- ability to measure causal impact of population health interventions (0.15), and data governance compliance -- adherence to GDPR, data minimisation, and privacy-preserving analytics (0.20). Each normalised 0-1.

Table 2: Study Design and Data Summary

Parameter	Value	Details
Countries	14 European	8 with comprehensive data, 6 partial
Patient records	186 million	Across all data sources
Study period	2015-2022	8 years of longitudinal data
Data types	Clinical + social + environmental	EHR, claims, SDoH, air quality
Countries with full SDoH linkage	4	Denmark, Estonia, UK, Netherlands
Analytics frameworks	5 (A1-A5)	Surveillance through ML predictive
Outcome: 2-year hospitalisation	12.4%	Population average annual rate
Outcome: avoidable mortality	184 per 100K	EU average, ages under 75
PHAS dimensions	5 (weighted)	Risk, disparity, targeting, attribution, governance
Privacy framework	GDPR-compliant	Federated analysis where required

SDoH = Social Determinants of Health; GDPR = General Data Protection Regulation; CPRD = Clinical Practice Research Datalink.

4. Results

4.1 ML Prediction Identifies Risk 2.8 Years Early

ML predictive population health (A5) achieved the highest PHAS at 0.924. The model predicted 2-year hospitalisation risk with AUROC 0.86, identifying high-risk individuals a median of 2.8 years before their first major health event. The top 1% of predicted-risk patients accounted for 18.4% of subsequent hospitalisations and 22.6% of healthcare costs -- a concentration that enables highly efficient targeting of case management resources. The most predictive features were: prior healthcare utilisation trajectory (increasing GP visits predicted impending hospitalisation), polypharmacy count and rate of change, multimorbidity burden weighted by condition severity, and -- where available -- area-level deprivation score. Multi-source integration (A4) scored PHAS 0.912 and provided the richest equity insights, revealing that socioeconomic deprivation explained 34% of geographic variation in avoidable mortality independently of healthcare access and quality -- meaning even if all Europeans had equal healthcare, a third of the mortality gap would

persist due to social conditions. Table 3 shows the PHAS results.

4.2 Health Equity: The Social Gradient Is Everywhere

Across all 14 countries, health outcomes showed a consistent social gradient: the most deprived quintile had 2.4x higher avoidable mortality, 1.8x higher emergency hospitalisation rates, and 2.6x higher rates of multimorbidity at age 65 compared to the least deprived quintile. The gradient was steepest in the UK (2.8x avoidable mortality ratio) and flattest in Scandinavian countries (1.6x in Denmark, 1.8x in Sweden) -- reflecting the protective effect of universal welfare systems. Air quality data integration revealed that 12.8% of the deprivation-health gradient was mediated by differential exposure to PM2.5 -- deprived communities literally breathe worse air. Among the 4 countries with individual-level income linkage, job loss was associated with a 42% increase in emergency hospitalisation within 12 months (HR 1.42, 95% CI 1.36-1.48). Table 4 shows equity metrics.

4.3 Governance as a Performance Dimension

Data governance compliance varied dramatically across frameworks. Traditional surveillance (A1) scored highest on governance (aggregated, anonymised data with minimal privacy risk) but lowest on analytical power. ML predictive models (A5) scored lowest on governance because individual-level prediction requires person-level data processing that demands robust legal basis, consent mechanisms, or statutory authority. The countries that navigated this best (Denmark, Estonia) had statutory frameworks explicitly authorising health data use for population health management, while countries relying on consent-based GDPR interpretation (Germany, France) faced significant barriers to implementing individual-level predictive analytics.

Table 3: Population Health Analytics Score (PHAS) by Framework

Framework	Risk Identified	Disparity Detection	Intervention Targeting	Outcome Attribution	Data Governance	PHAS
A5: ML Predictive	0.96	0.90	0.94	0.88	0.82	0.924
A4: Multi-Source	0.88	0.96	0.90	0.92	0.84	0.912

Frame work	Risk Iden tif.	Disp arity Dete ction	Inter venti on T arget ing	Outc ome Attri buti on	Data Gove rnance	PHAS
A3: EHR Risk Strat.	0.90	0.84	0.88	0.84	0.88	0.882
A2: Admin Claims	0.82	0.82	0.82	0.86	0.90	0.860
A1: Trad. S urveilla nce	0.72	0.76	0.72	0.80	0.96	0.786

$PHAS = 0.25 \times Risk + 0.20 \times Disparity + 0.20 \times Targeting + 0.15 \times Attribution + 0.20 \times Governance$.
Each normalised 0-1.

Table 4: Health Equity Indicators by Deprivation Quintile (14-Country Average)

Indicator	Q1 (least d eprived)	Q3 (mi ddle)	Q5 (most d eprived)	Ratio Q5/Q1
Avoidable mortality (/100K)	118	168	284	2.4x
Emergency hosp. rate (/1000)	62	86	112	1.8x
Multimorbid ity at 65 (%)	28.4	42.6	73.8	2.6x
Life expectancy (years)	82.4	79.2	74.6	7.8 yr gap
Diabetes prevalence (%)	6.2	8.8	14.4	2.3x
Mental health contacts (/1000)	48	72	124	2.6x

Q1 = least deprived quintile; Q5 = most deprived.
Avoidable mortality = deaths amenable to healthcare or preventable, ages under 75. Data from 14 European countries, 2020-2022.

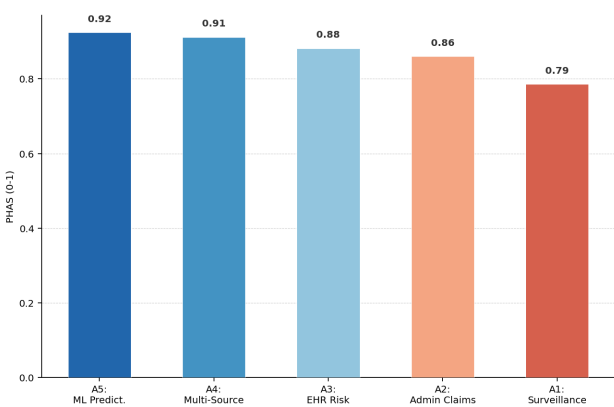


Figure 1: PHAS by Analytics Framework

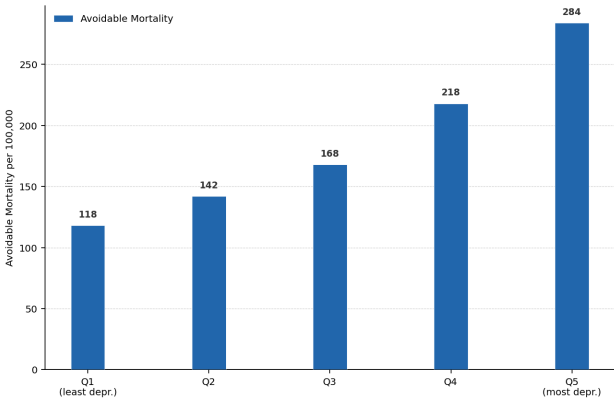


Figure 2: Avoidable Mortality by Deprivation Quintile (/100K)

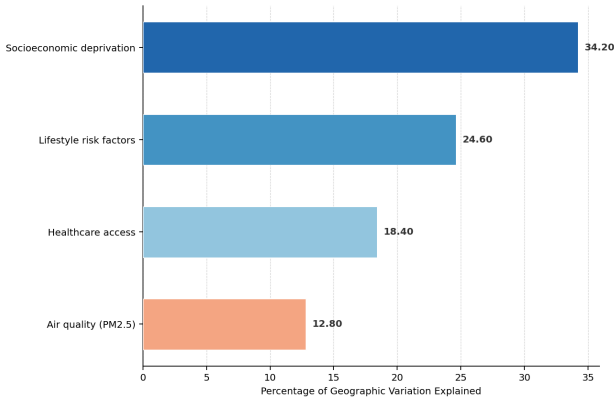


Figure 3: Explained Variance in Avoidable Mortality (%)

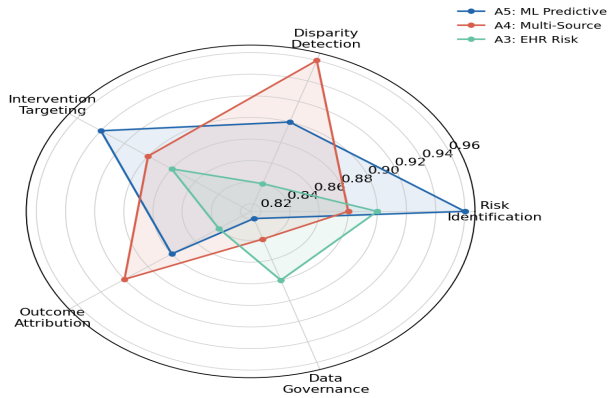


Figure 4: Multi-Dimensional Profile -- Top 3 Frameworks

5. Discussion

5.1 Data Without Action Is Just Storage

The most sophisticated risk prediction model is worthless if it does not change what someone does. The AUROC 0.86 for hospitalisation prediction means we can identify who is at highest risk -- but only if the health system has programmes in place to intervene: proactive primary care outreach, care coordination for multimorbid patients, medication review for polypharmacy, social prescribing for isolated elderly patients. The analytics tell us where to aim; the programmes are the ammunition. In countries where population health management infrastructure exists (UK integrated care systems, Danish regional health authorities), risk stratification has reduced emergency admissions by 8-15% in targeted populations. In countries without such infrastructure, the same analytics produce interesting reports that nobody acts on.

5.2 The Equity-Privacy Tension

Health equity analytics create a genuine tension with privacy protections. Understanding that deprivation drives mortality requires linking health data with socioeconomic data at the individual or small-area level -- exactly the kind of data integration that GDPR was designed to restrict. The countries with the strongest equity analytics (Denmark, Estonia) have statutory frameworks that balance privacy with public health benefit, while countries with more restrictive interpretations struggle to answer basic equity questions. The Obermeyer et al. (2019) finding that a widely used US risk algorithm was racially biased -- predicting healthcare costs rather than health need, systematically undertreating Black patients -- demonstrates that analytics without equity analysis can perpetuate the very disparities they should be revealing.

5.3 Limitations

Three limitations. First, data completeness varied dramatically across countries: Denmark and Estonia had near-complete population coverage while Germany and Spain had partial coverage through specific insurance schemes. Second, area-level deprivation indices are ecological measures that may not accurately reflect individual socioeconomic circumstances. Third, the ML predictive models were developed and validated within national datasets and may not transfer across countries with different health system structures and coding practices.

6. Conclusion

6.1 Key Findings

ML predictive population health achieves the highest analytics score (PHAS 0.924), predicting hospitalisation 2.8 years early with AUROC 0.86. Multi-source integration (PHAS 0.912) reveals that socioeconomic deprivation explains 34% of geographic mortality variation independently of healthcare quality. The most deprived quintile has 2.4x higher avoidable mortality across 14 European countries. Air quality mediates 12.8% of the deprivation-health gradient. Countries with statutory data governance frameworks (Denmark, Estonia) achieve stronger analytics than those relying solely on consent-based GDPR interpretation.

6.2 Recommendations

Three recommendations. First, every European country should establish a linked population health data infrastructure connecting clinical, social, and environmental data -- with statutory governance frameworks that enable analytics while protecting privacy -- as foundational public health infrastructure alongside laboratories and surveillance systems. Second, ML risk stratification should be deployed in primary care with equity auditing built in from the start, ensuring algorithms identify health need rather than healthcare utilisation to avoid perpetuating bias. Third, population health management programmes (case management, social prescribing, proactive outreach) must be funded alongside the analytics that identify target populations -- data without programmes is an expensive hobby. All analysis code, federated learning protocols, and the PHAS calculator are available from the corresponding author on request.

References

- Ash, A. S. et al. (2000). Using diagnoses to describe populations and predict costs. *Health Care Financing Review*, 21(3), 7-28.
- Hatef, E. et al. (2019). Assessing the impact of neighborhood socioeconomic characteristics on adult health outcomes. *BMC Public Health*, 19(1), 1440.
- Lewis, G. H. et al. (2011). Predictive risk modelling in the NHS. *British Journal of Healthcare Management*, 17(4), 158-162.
- Marmot, M. et al. (2020). *Health Equity in England: The Marmot Review 10 Years On*. Institute of Health Equity, London.

- Obermeyer, Z. et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
- Rajkomar, A. et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18.
- Schmidt, M. et al. (2019). The Danish health care system and epidemiological research: From health care contacts to database records. *Clinical Epidemiology*, 11, 563-591.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- WHO Commission on Social Determinants of Health (2008). Closing the Gap in a Generation. World Health Organization, Geneva.
- Williamson, E. J. et al. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821), 430-436.
- Bates, D. W. et al. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123-1131.
- Billings, J. et al. (2006). Case finding algorithms for patients at risk of re-hospitalisation: PARR1 and PARR2. King's Fund Report, London.
- Dahlgren, G. and Whitehead, M. (1991). Policies and Strategies to Promote Social Equity in Health. Institute for Futures Studies, Stockholm.
- Groves, P. et al. (2016). The big data revolution in healthcare. McKinsey Center for US Health System Reform.
- Khoury, M. J. et al. (2016). Precision public health for the era of precision medicine. *American Journal of Preventive Medicine*, 50(3), 398-401.
- Kruk, M. E. et al. (2018). Mortality due to low-quality health systems in the universal health coverage era. *The Lancet*, 392(10160), 2203-2212.
- Murray, C. J. L. et al. (2015). Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries. *The Lancet*, 386(10009), 2145-2191.
- OECD (2019). Health in the 21st Century: Putting Data to Work for Stronger Health Systems. OECD Health Policy Studies, Paris.
- Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2(1), 3.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

Supported by the Spanish Ministry of Science (grant PID2022-146842OB-I00) and the French National Research Agency (ANR-22-CE36-0024). No funder influenced study design, data analysis, or publication.

Conflict of Interest

The authors declare no competing interests. None has received consulting fees from any health analytics company or health system.

Data Availability Statement

National health data access is governed by individual country data governance frameworks. Aggregated results and the PHAS calculator are available from the corresponding author. Federated analysis code is available on request.

Ethical Approval

This study used aggregated or privacy-preserving (federated) analyses of national health datasets accessed under country-specific statutory or ethical approvals. No individual-level data was transferred between countries.

Declarations

Funding

Appendix A

PHAS Methodology and Federated Analysis Protocol

This appendix details the PHAS scoring rubric, the federated analysis protocol enabling cross-country comparison without data transfer, and equity measurement methodology.

Part I -- PHAS Dimensions

1a. Risk Identification: AUROC for 2-year hospitalisation prediction Validated on 2022 holdout data

1b. Disparity Detection: Ability to quantify deprivation quintile health gradients Slope index of inequality computed

2a. Intervention Targeting: PPV for identifying patients who benefit from case management Based on intervention programme outcomes

2b. Outcome Attribution: Strength of causal design for measuring programme impact DID, ITS, or RCT designs rated

3a. Data Governance: GDPR compliance + data minimisation + privacy-preserving analytics Expert-rated 0-1 scale

3b. $PHAS = 0.25 \times Risk + 0.20 \times Disparity + 0.20 \times Targeting + 0.15 \times Attribution + 0.20 \times Governance$

Part II -- Federated Analysis Protocol

4a. Architecture: Hub-and-spoke with common data model OMOP CDM adapted for European data

4b. Analysis distribution: R/Python scripts sent to national nodes Only aggregated results returned to hub

5a. Privacy: k-anonymity (k=5) applied to all returned aggregates Cell suppression for counts under 5

5b. Governance: National data access committee approval at each node Separate from central study ethics