



Harmonizing Johannesburg Cohorts for Climate–Health Analytics

Reproducible multi-source data integration for machine-learning heat-vulnerability research



Benjamin Devantier-Du Plessis^{1,3}, Craig Parker^{1*}, Sam Zermatten-Jack¹, Natasha Lalloo¹, Nicholas B. Brink¹, Christopher Jack², Matthew Chersich^{3,1}, Guéladio Cissé^{4,5}, Sibusisiwe Makhanya⁶, and the HE²AT Centre Consortium

¹ Wits Planetary Health Research · ² CSAG, University of Cape Town · ³ Trinity College Dublin · ⁴ Swiss TPH · ⁵ UPGC · ⁶ IBM Research Africa

An ALCOA+-compliant¹⁵ analysis dataset linking 17 Johannesburg cohort studies (2003–2021) to high-resolution climate exposure and ward-level socioeconomic context.

17

cohort studies

11,114

clinical visits

9,756

unique patients

27

biomarkers · 8 systems

5

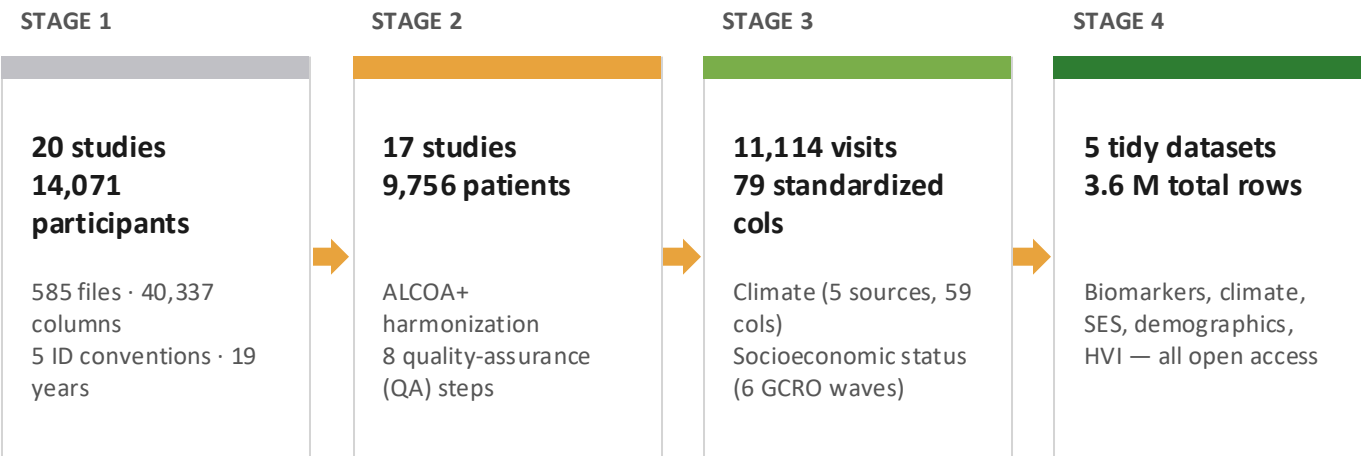
tidy long-format datasets

Background

Johannesburg holds 20+ years of clinical research — but every study uses its own IDs, units, files and dates, blocking climate-health analysis at scale.

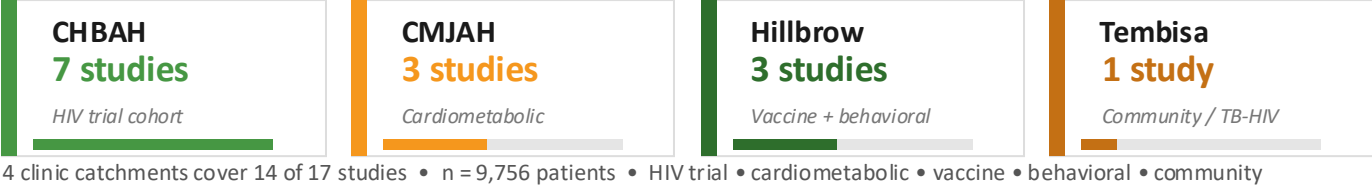
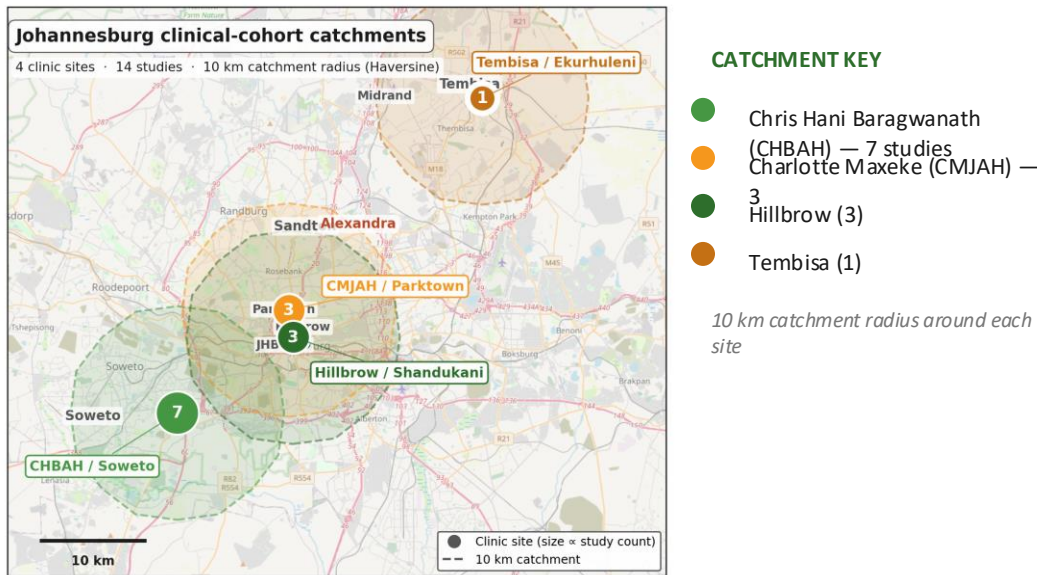
We present an end-to-end pipeline that harmonizes 17 JHB cohort studies (from 20 candidates) into one analysis-ready dataset linked at every visit to high-resolution climate, ward-level socioeconomics and the JHB Heat Vulnerability Index — **reproducible, auditable, climate-aware**.

The Harmonization Funnel



Pre-harmonization chaos → ALCOA+-compliant analysis dataset → reproducible long-format inputs

Clinic Catchments



Harmonization Methods

ALCOA+ Data-Integrity Framework — 8 QA steps, 2-round audit log

OUTPUTS — 27 biomarkers across 8 organ systems + 5 tidy long-format datasets released							
Cardiovascular	Lipid	Renal	Hepatic	Metabolic	Body comp.	Inflammatory	Immunological
systolic BP diastolic BP · HR	fasting glucose HbA1c · insulin	systolic BP diastolic BP · HR	ALT · AST	fasting glucose HbA1c · insulin	ALT · AST	hs-CRP (ESR unavailable)	CD4 · viral load (CD8 unavailable)
TIDY clinical biomarkers		TIDY climate variables		TIDY demographics		TIDY socioeconomic	
650 K rows 27 biomarkers		2.8 M rows 59 climate cols		36 K rows age · sex · HIV · BMI		40 K rows GCRO-imputed SES	
						40 K rows IPCC HVI / visit	

Methodology is portable — Abidjan extension in planning (8 PAC-CI studies, 76 000 records) · primary key (study_source, patient_id, visit_date) links all datasets.

Corresponding
Craig Parker · craig.parker@witsphr.org · Wits PHR

Acknowledgements
Abbreviations ALCOA+ Attributable, Legible, Contemporaneous, Original, Accurate (FDA data-integrity standard) · GCRO Gauteng City-Region Observatory · HVI Heat Vulnerability Index · HREC Human Research Ethics Committee · MDB Municipal Demarcation Board · ECMWF European Centre for Medium-Range Weather Forecasts · LP DAAC Land Processes Distributed Active Archive Center.
Funding: HE²AT Center is supported by the NIH Common Fund, managed by the Fogarty International Centre (RIC), National Institute of Environmental Health Sciences (NIEHS), and OD/Office of Strategic Coordination (OSC) of the National Institutes of Health (NIH award number: U54TW012083). Disclaimer: The poster content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Wits HREC ref. 200606. Data partners: GCRO · MDB Wards 2020 · Copernicus / ECMWF · NASA LP DAAC. Authors at Wits hold indirect fossil-fuel investments via institutional pension funds.

Climate Linkage

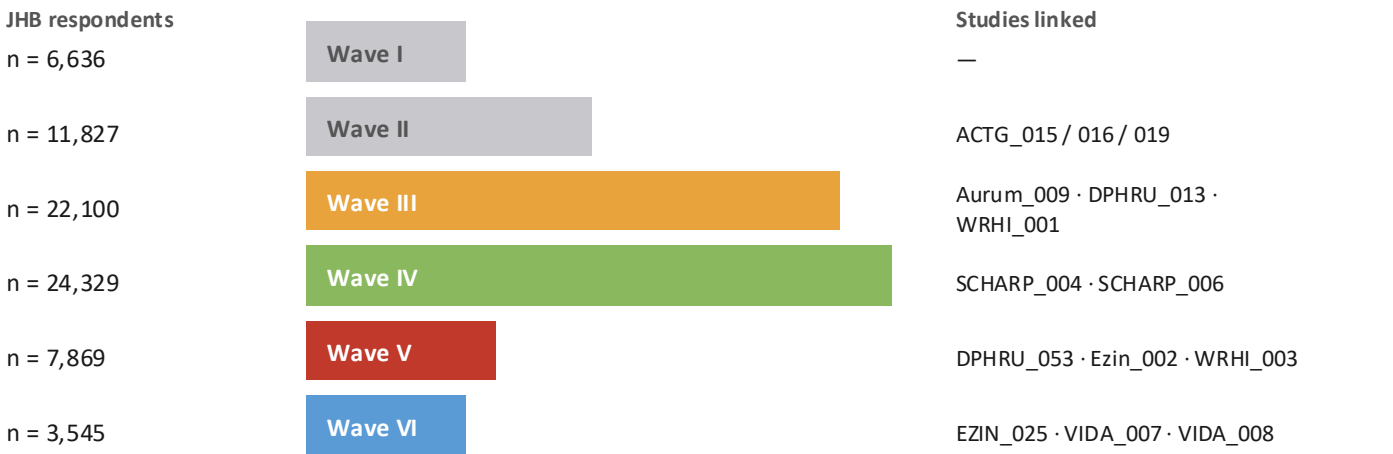
Each visit linked spatiotemporally to 5 datasets · 59 engineered columns

Copernicus / ECMWF ERA5	Copernicus / ECMWF ERA5-Land	NASA LP DAAC MODIS MOD11A1
● Direct measurements (1–4)		● Derived indices (5–10)
1 Same-day temperature mean · max · min · diurnal range		6 Humidity & wet-bulb globe T (WBGT) Magnus · Stull wet-bulb · ISO 7243
2 Rolling-mean lags 1, 3, 7, 14, 30 d (machine-learning inputs)		7 90th-pctl heatwave Perkins–Alexander 2013
3 Variability 7-d / 30-d std · centered anomaly		8 Excess Heat Factor Nairn–Fawcett 2015 (4-tier)
4 Surface temperature ERA5 land-surface T (LST) mean & max		9 Seasonal context departure · seasonal percentile
5 Wind & thermal indices Steadman apparent T (AT) · Rothfusz heat index (HI)		10 Distributed-lag non-linear model (DLNM) lags 0–21 (Gasparrini 2011)

59 engineered climate columns per visit · all peer-reviewed methods^{1–8,10} · auto-fallback for ERA5 corruption

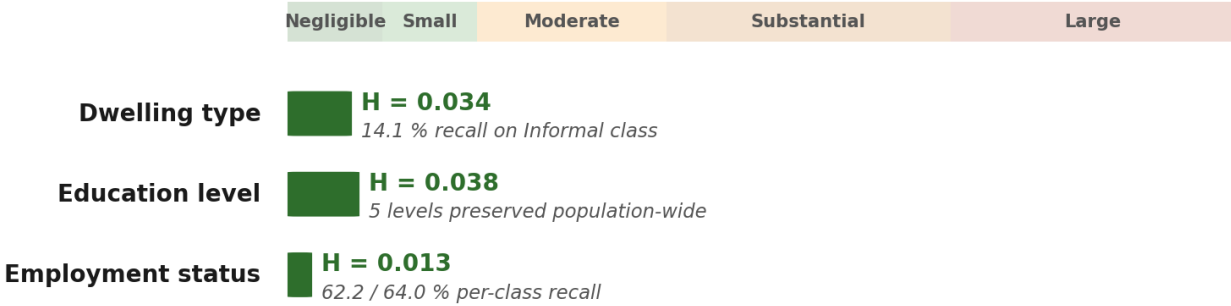
Socioeconomic Linkage

K = 20 ward-based nearest-neighbour fusion from 6 GCRO Quality of Life waves^{9–10}
· donors within **10 km** of clinic (Haversine, MDB Wards 2020) · exact-match sex, ±5-yr age caliper
· distance-weighted draw **w = 1/(d+1)** · income = weighted median · seed = 42 (cross-seed deviation < 6 %)
IPCC HVI mapping: dwelling → **Exposure** · income → **Sensitivity** · education / employment → **Adaptive Capacity**^{17–19}



Validation: distributional fidelity (Hellinger H, 20 cross-validation draws) — all SES proxies recover the population distribution within negligible-to-small distanceCRO Quality of Life Survey wave year (2009 → 2020)

Aggregate cross-validation distributional fidelity



Scan for code & data

References & resources
Climate methods [1–8] Hersbach H et al. (2020) ERA5. QJRM 146:1999. · Wan Z et al. (2015) MOD11A1 V006. NASA LP DAAC. · Steadman R (1984) Apparent T. JCAM 23. · Rothfusz L (1990) NWS SR 90-23. · Stull R (2011) JAMC 50:2267. · Perkins SE & Alexander LV (2013) J Climate 26:4500. · Nairn JR & Fawcett RJB (2015) IJERPH 12:227. · Gasparrini A (2011) dlmm, JSS 43(8).
SES data fusion [9–14] D'Orazio M et al. (2006) Statistical Matching: Theory & Practice. Wiley. · Rässler S (2002) Statistical Matching. Springer. · D'Orazio M (2024) Statistical Papers 65:1547–1581. · Conti PL et al. (2017) J Off Stat 33:541–571. · Diez Roux AV (2001) Am J Public Health 91:1783–1789. · Andridge RR & Little RJA (2010) Int Stat Rev 78:40–64.
Data integrity [15–16] FDA (2018) Data Integrity & Compliance with Drug CGMP, Q&A Guidance. · GCRO Quality of Life Survey (2020/21) — gcro.ac.za/outputs/qol-survey. · JHB HVI [17–19] Parker C et al. (2025) Int J Biometeorol 69:2501. · Cutter SL et al. (2003) Soc Sci Q 84:242–261. · Reid CE et al. (2009) Environ Health Perspect 117:1730–1736.

