International Journal of Epidemiology, 2026, **55(4)**, dyag124
https://doi.org/10.1093/ije/dyag124
Data Resource Profile

IEA International Epidemiological Association
OXFORD

Data Resource Profile: Multimorbidity in Africa Digital Innovation, Visualisation and Application (MADIVA) research hub

Daniel Ohene-Kwofie[1,2,*], Nkosinathi Masilela[1], Kerry Glover[3], Samuel Iddi[4,5], Daniel Maina[4], Molulaqhooa Linda Maoyi[6], Theophilous Mathema[3], Mwiza Gideon Singini[7], Scott Hazelhurst[2,3], Chodziwadziwa Whiteson Kabudula[1] as members of the MADIVA Research Hub and the DS-I Africa Consortium[†]

[1]SAMRC/Wits Rural Public Health and Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa
[3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[4]African Population and Health Research Center (APHRC), APHRC Campus, Nairobi, Kenya
[5]Department of Statistics and Actuarial Science, University of Ghana, Legon, Accra, Ghana
[6]DSTI-SAMRC South African Population Research Infrastructure Network (SAPRIN), South African Medical Research Council, Durban, South Africa
[7]Non-Communicable Diseases Research Unit (NCDRU), South African Medical Research Council, Cape Town, South Africa
*Corresponding author. SAMRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of the Witwatersrand, School of Public Health, 27 St Andrews Road, Parktown 2193, South Africa. E-mail: daniel.ohene-kwofie@wits.ac.za
[†]See Acknowledgment for a complete list of authors who are the members of the MADIVA Research Hub and the DS-I Africa Consortium.

Keywords multimorbidity, sub-Saharan Africa, data harmonization, health and demographic surveillance system, data science, non-communicable diseases

Key Features

- The Multimorbidity in Africa Digital Innovation, Visualisation, and Application (MADIVA) research hub was established to facilitate the study of multimorbidity and associated risk factors, disease clustering, and stratification in African populations.
- The hub has harmonized and integrated data from population-based health and demographic surveillance system (HDSS) platforms from two sub-Saharan African settings representing both rural and urban populations.
- The platform consists of baseline population data collected from 1992 to 2024 in Agincourt, South Africa and from 2002 to 2019 in Nairobi, Kenya, with annual follow-ups and 10 nested health-related cohorts.
- The HDSS platforms contribute, overall, a total of >4 million person-years of observation from 574 598 individuals (54.8% females, 45.2% males). The harmonized nested cohorts contribute 130 550 observations on non-communicable diseases from 67 478 individuals (58.7% females, 41.3% males; mean age 44.5 years; 19.34 SD).
- The main categories of the data include socio-demographic characteristics, household and individual socio-economic status, and self-reported and objectively measured histories of chronic diseases and their treatments, as well as lifestyle behaviours. In addition, genotyping data are available for ∼6000 individuals.
- Data from the MADIVA hub provide a valuable resource for the longitudinal analyses of multimorbidity patterns, exploration of disease clustering, and the development of context-specific public health interventions.
- Interested researchers can contact the hub at info@madiva.africa or visit https://www.madiva.africa.

Data resource basics

The burden of non-communicable diseases (NCDs) is rapidly increasing across Africa, driven by demographic and lifestyle transitions [1]. Concomitantly, levels of multimorbidity—commonly defined as the coexistence of two or more chronic conditions in an individual—are also rising [2, 3]. Multimorbidity is associated with increased premature mortality [4], lowered quality of life [5], diminished mental health [6, 7], increased risk of polypharmacy [8], and intensified health services utilization and associated costs, particularly in resource-poor settings [9]. Despite its significant impact, multimorbidity remains relatively

Received: 17 November 2025. **Accepted:** 25 June 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of the International Epidemiological Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

under-researched compared with individual disease conditions [10].

Despite the complexities associated with multimorbidity and its rising burden in sub-Saharan Africa, most of the studies conducted have been in silos [11], making comparison of the findings difficult across the region. The systematic harmonization and integration of multiple datasets for multimorbidity may increase sample sizes and the comparability of study findings [12–14]. Additionally, data collected from longitudinal studies as well as Health and Socio-demographic Surveillance sites are a good source for understanding the dynamics of diseases in a population [15]. While individual datasets from various sources often capture similar constructs by using different measures, data harmonization enables comparability across these heterogeneous sources and facilitates their integration in a consistent and coherent manner. This facilitates cross-study comparisons to monitor, measure, and verify evolving population-level trends. In this data resource profile, we provide an overview of a novel data resource for the study of multimorbidity and associated risk factors in Africa produced by the Multimorbidity in Africa Digital Innovation, Visualisation and Application (MADIVA) research hub.

MADIVA is a research hub of the Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative. The goal of DS-I Africa is to leverage data science technologies to transform biomedical and behavioural research, including developing solutions that lead to improved health for individuals and populations in Africa. As part of MADIVA's objectives, the hub is utilizing secondary data from two health and demographic surveillance system (HDSS) sites—rural Bushbuckridge (Agincourt), South Africa, and two urban informal settlements (Korogocho and Viwandani) of Nairobi, Kenya collected by the African Population and Health Research Center (APHRC)—to produce a harmonized data resource for the study of multimorbidity in Africa. Both sites have rich datasets from longitudinal and cohort studies gathered by the HDSS as well as nested research studies spanning several years. These existing data resources include demographic data, clinical health records, and genomic data, which are valuable for the study of multimorbidity. The aim of this paper is to discuss the data resource, with a particular emphasis on the analytical dataset created from multiple secondary data sources. It provides a comprehensive foundation for examining different health conditions, multimorbidity, and the associated health effects, including comprehensive studies of illnesses trends, risks factors, and long-term consequences across diverse populations.

Data collected

The data included in the MADIVA data resource come from two primary sites. The first is the Agincourt HDSS (hereafter Agincourt), which is in the rural north-east of South Africa, close to the border with Mozambique. Agincourt was started in 1992 and currently includes an active population of ∼117 000 individuals from ∼22 000 households in 31 villages [16]. The second is the Nairobi Urban Health and Demographic Surveillance System (NUHDSS), which was established in 2002 in two slum communities in Nairobi (Korogocho and Viwandani). The NUHDSS platform investigates the long-term consequences of urban slum residence on health and socio-economic outcomes and, as of 2019, consisted of ∼88 974 active individuals from ∼33 462 households [17].

The current MADIVA database comprises representative, population-based, individual-level longitudinal and health surveys, including biomarkers, conducted in 2010 or later (for the Agincourt nested studies) and 2007 or later (for the NUHDSS nested studies). The research hub leveraged six studies (five longitudinal and one observational) from Agincourt and four longitudinal studies from the NUHDSS. Although all these nested studies were conducted independently at the respective sites at different time points, they all include biomarker information as well as self-reported responses that provide valuable information for the study of multimorbidity. In addition to the studies nested in the HDSS sites, the harmonization included socio-demographic and household information as well as individual event history from the respective HDSS platforms. Table 1 provides an overview of the various studies incorporated into the MADIVA data-harmonization process, detailing the data time points, volume, and types of data collected.

Datasets from the indicated studies (see Table 1) and the HDSS were collated into a relational database and then harmonized by using Structured Query Language. The harmonization process was guided by a stepwise and iterative approach, involving data-specification definitions and approval. The data-specification document highlights the expected variables, variable naming conventions, and the value domains, such as 0—no, 1—yes, 333—don't know, 444—refused, 888—not applicable, 999—missing, etc. The current version of the data-specification document is available at http://github.com/MADIVA-DSI/data. Based on the data specification, we documented the variables collected for each study, including identifying domains and core target variables for each domain relevant for the study of multimorbidity.

The dataset from each integrated study was then extracted based on the data specification. This complex process involved aligning variable definitions across the studies (i.e. identifying and reviewing which questions align across studies as per the data specification), checking data quality and ensuring consistency, especially for biomarker measures, documenting variable skip patterns, and confirming which responses were missing from respective studies.

After the extraction, the dataset went through a rigorous review, validation, and profiling process to identify any data-quality issues from the extraction process. Data quality was ensured through a structured, multilayered approach involving study-specific internal consistency checks, as well as review and audits of harmonized datasets. Initial cleaning addressed duplicates, implausible values, coding non-responses and missing values, and logical inconsistencies—particularly in core demographic and temporal variables such as age, sex, and event dates (e.g. birth, death, and diagnosis). The data-profiling stage provided descriptive statistics with graphical representations for each variable enabling seamless review such as range checks, frequency distributions, and cross-tabulations to detect anomalies and verify the integrity of transformed and integrated variables. To assess completeness, we also calculated variable-level missingness across sites and time points, and visualized patterns to identify potential biases. Where necessary, the original metadata, data dictionaries, and survey instruments were

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

Table 1 Detailed description of the included projects providing data for the MADIVA harmonization and integration.

Project	Data collected	Study design	Countries	Data volume	Study period	Age (years) [mean (SD)]
Health and socio-demographic platform [16, 17, 23]	Demographic and social factors, health behaviours, medical history and exposures Supplementary Table 1 presents the detailed person-years and mortality rates by year for each of the sites	Longitudinal	South Africa and Kenya	South Africa: 35 912 households, 311 633 individuals, 31 villages Kenya: 84 903 households, 262 915 individuals	South Africa: annually since 1992 Kenya: 2 or 3 rounds per year from 2002 to 2019	All ages
Verbal autopsy (VA) [24]	The World health Organization (WHO)-based VA questionnaire providing data on symptoms of the illness or injury prior to death, including health-seeking behaviour for all deceased individuals to facilitate automated cause-of-death assignment Supplementary figure S1a and b provide broad causes of death for each HDSS site	Cross-sectional	South Africa and Kenya	South Africa: 21 270 deceased individuals Kenya: 5474 deceased individuals	South Africa: annually since 1992 Kenya: 2002–17	All ages: 48.7 (25.56); 32.2 (21.17)
Africa Wits-INDEPTH study of genomic and environmental factors for cardiometabolic risk (AWI-Gen 1,2) [25, 26]	Socio-demographic characteristics, lifestyle behaviour, medical history of chronic and infectious diseases, anthropometry, body fat distribution, blood pressure, and blood and urine biomarkers Genomic data	Longitudinal	South Africa and Kenya	South Africa: Wave 1 = 2486; Wave 2 = 1255 Kenya: Wave 1 = 2003; Wave 2 = 1189 4489 genotyped 30 whole-genome sequences	South Africa: 2014–15; 2019–21 Kenya: 2014–15; 2020–21	38+ years: 58.7 (11.03); 62.6 (10.54); 48.6 (5.85); 54.0 (5.74)
African Research on Kidney Disease (ARK) Study Phase 1 and 2 [27]	Socio-demographic, anthropometric, and clinical data; blood samples for creatinine, hepatitis B serology; spot urine for dipstick testing and urine albumin: point-of-care: HIV infection, diabetes, and hypercholesterolaemia Genomic data	Longitudinal	South Africa	South Africa: Wave 1 = 2759; Wave 2 = 1505 630 genotyped	South Africa: 2017–18 ; 2018–19	20–79 years: 37.2 (13.86); 41.2 (14.58)

(continued)

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

Table 1 (continued)

Project	Data collected	Study design	Countries	Data volume	Study period	Age (years) [mean (SD)]
Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) 1, 2, 3 [28]	Demographic and social factors, health behaviours, anthropometric measurements, biomarkers, cognition, mental health, and self-reported health Genomic data	Longitudinal	South Africa	South Africa: Wave 1 = 5059; Wave 2 = 4176; Wave 3 = 3707 4000+ genotyped	South Africa: 2014–15; 2018–19; 2021–22	40+ years: 62.3 (13.00); 65.1 (12.57); 66.6 (12.00)
HIV/cardiometabolic risk factor survey—HIV NCD [29]	Sexual behaviour and chronic-disease risk factors; anthropometric measurements; and biomarkers for diabetes, cholesterol, and dried blood spots (DBS) for HIV	Cross-sectional	South Africa	4704 individuals	2010–11	15+ years: 42.0 (19.23)
Nkateko Trial—Hypertension Surveillance 1 and 2 [30]	Clinic-based chronic-disease data: HIV, hypertension, diabetes; physical measurements (blood pressure)	Longitudinal	South Africa	Wave 1 = 3772; Wave 2 = 3015	2013–14 ; 2014–15	17+ years: 56.2 (19.26); 52.3 (19.18)
Individual Health 1 and 2	Nested within the HDSS platform, collected anthropometric measurement (BMI), HIV status (rapid test + DBS), self-reported general health (HIV, diabetes, hypertension, TB)	Longitudinal	South Africa	Wave 1 = 36 809; Wave 2 = 37 278	2021–22; 2022–23	18+ years: 38.1 (18.82); 39.6 (19.65)
Assessing the links between socio-economic status, perceived personal risk, and risk factors for cardiovascular and related NCDs in a population of slum dwellers in Nairobi, Kenya [31]	Assessed the connections between socio-economic status (SES), perceived personal risk, and risk factors for cardiovascular and other NCDs	Longitudinal	Kenya	Alcohol = 106; main = 5396	2007–09; 2008–09	18+ years: 38.8 (15.80); 43.2 (14.89)

(continued)

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

Table 1 (continued)

Project	Data collected	Study design	Countries	Data volume	Study period	Age (years) [mean (SD)]
A community-based intervention for primary prevention of cardiovascular diseases in the slums of Nairobi: the SCALE-UP (SU) study protocol for a prospective quasi-experimental community-based trial (SU Population Baseline & Endline, SU follow-up clinic) [32]	Demographic and socio-economic data; anthropometric and clinical measurements including CVD behavioural risk factors, physical measurements (weight, height, waist and hip circumference, blood pressure), glucose	Longitudinal	Kenya	First clinic = 721; follow-up clinic = 3313; clinic baseline = 604; clinic end line = 509; population baseline = 5649; population end line=4466	2011–14; 2012–14 ; 2012–14; 2013–14; 2012–13; 2013–14	35+ years: 51.3 (17.89); 54.0 (15.29); 51.3 (17.38); 52.1 (16.77); 45.5 (12.78); 46.8 (9.35)
Diabetic (baseline and end line) [33]	Socio-demographic and clinical characteristics of study participants including health behaviour—diet (salt intake) and exercise, healthcare-seeking, as well as blood pressure, anthropometric measurements—BMI, waist, hip; fasting blood glucose	Longitudinal	Kenya	Baseline = 85; end line =47	2009–10; 2010–11	18+ years: 53.9 (15.58); 54.8 (12.89)

BMI, body mass index; CVD, cardiovascular disease; TB, tuberculosis; DBS, dried blood spots; WHO, World Health Organization.

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

consulted to support variable interpretation, with direct input from local data managers to resolve context-specific discrepancies. These quality-control procedures were repeated iteratively during harmonization, enhancing the reliability and comparability of the integrated datasets.

The finalized datasets in the relational database were imported into STATA, where the variables were carefully documented based on the data specifications. The resulting documented STATA dataset was used to generate comprehensive metadata documentation based on the Data Documentation Initiative (DDI) [18] standard and the international standard for data and metadata exchange using the DDI Metadata Editor (Nesstar Publisher [19]). The current version of the MADIVA DDI documentation is available at the repo http://github.com/MADIVA-DSI/data. The integration and harmonization processes are summarized in Fig. 1. A profile of the main categories of datasets generated from the integration and harmonization process is presented in Table 2.

The NCD indicators dataset

To facilitate the study of multimorbidity in Africa, the NCD indicators dataset leveraged data from existing health-related nested studies, which had their own research objectives but shared a common interest in NCDs, human immunodeficiency virus (HIV), and biomarkers, and their relationship with socio-demographic factors of the surveillance populations covered. The dataset was collated by the Data Management and Analytics Core of the MADIVA research hub, to meet the need for valid, reliable, and well-curated and well-documented data to help to understand, characterize the burden, and address the complexities associated with multimorbidity in Africa. Currently, 67 478 (∼12%) of the combined HDSS unique individuals have at least one observation from these NCD and health surveys, which were nested within the respective HDSS sites. The NCD and health surveys contribute a total of 130 550 observations, mainly among individuals aged ≥20 years, with close to 59% being observations from females. The participants included residents of NUHDSS (with 11 864 unique individuals who contributed 23 965 observations) and Agincourt (with 55 614 unique individuals who contributed 106 585 observations).

The variables harmonized and integrated from these surveys are broadly categorized into the following domains:

i) socio-demographic: sex, age, marital status, educational level, employment status;
ii) health examinations and measurements such as blood pressure, pulse, height and weight, body mass index (BMI), waist and hip circumference;
iii) physical-activity status: activities of daily living, vigorous, moderate exercise, sedentary activity, pain;
iv) self-reported health conditions: diagnosis and treatment (including traditional treatment) of HIV, hypertension, diabetes, tuberculosis (TB), stroke, angina, high cholesterol, kidney disease, asthma, cancers, thyroid disease;
v) healthcare utilization: number of health-facility/clinic visits;
vi) behavioural risk factors and health behaviours: tobacco use, alcohol consumption, diet (fruit, vegetables, soft drinks);
vii) sexual behaviour: condom use, number and types of sexual partners;
viii) point-of-care and laboratory results: blood assays: total cholesterol, high-density lipoprotein, low density lipoprotein, triglycerides, glucose, haemoglobin, HIV (viral load, antiretroviral therapy).

More information describing the data categories and variables can be found on page 4 of the Supplementary Material, available as Supplementary data at *IJE* online.

Data resource use

The harmonized MADIVA dataset has only recently been assembled, making it a new and promising resource for studying NCDs and multimorbidity within HDSS populations across both rural and urban settings in sub-Saharan Africa. Although there have been a few publications based on the harmonized dataset [20–22], several analyses are currently underway. Initial work focuses on examining the prevalence and patterns of multimorbidity across the surveillance populations and assessing the trajectories of major chronic conditions and cardiometabolic risk factors contributing to multimorbidity, including HIV, hypertension, diabetes, stroke, and obesity.

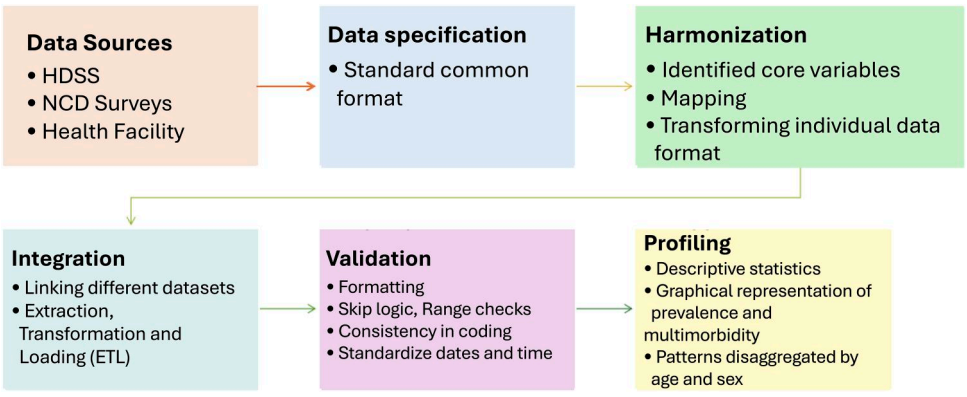


Figure 1 MADIVA data-harmonization and data-integration process. ETL, Extract, Transform, Load.

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

Table 2 Profile of the data categories generated as part of the MADIVA database highlighting the data volume by sex for each HDSS site.

Dataset	Description	Variable count	Number of records (*N*)	Agincourt HDSS, South Africa			NUHDSS, Nairobi		
				Male (%)	Female (%)	*N* (%)	Male (%)	Female (%)	*N* (%)
Individuals	This dataset comprises demographic information on all unique individuals from the HDSS, including when they entered the surveillance area and when they moved out of the surveillance area	26	574 598 (∼4 million person-years)	139 428 (44.73)	172 243 (55.26)	311 677 (54.24)	120 520 (45.84)	142 395 (54.16)	262 921 (45.76)
Individual events	Provides the detailed event history observations of all individuals within the HDSS, including when and how they entered the surveillance (either through enumeration, birth or in-migration), as well as their current status within the HDSS (i.e. still current within the HDSS, moved out, or died)	7	164 5535	287 370 (44.55)	357 638 (55.45)	630 042 (38.28)	466 470 (45.94)	549 023 (54.06)	1 015 493 (61.71)
Individual socio-economic (SES) indicators	The dataset provides observations on the education, employment and marital status of individuals within the HDSS, captured across surveillance rounds	6	681 1985	1 179 931 (47.81)	1 287 988 (52.19)	2 464 207 (36.17)	2 043 496 (46.67)	2 334 860 (53.33)	4 347 778 (63.83)
Household memberships	The dataset consists of the membership observation records of everyone within the HDSS, linking individuals to their household and indicating their relationship with the head of the household for each episode	8	733 640	160 902 (44.94)	197 133 (55.06)	358 035 (48.80)	170 303 (45.34)	205 302 (54.66)	375 605 (51.20)
Households	Households describes the socio group for the individuals within the surveillance	4	120 815	N/A[1]	N/A	35 912 (29.72)	N/A	N/A	84 903 (70.28)
Household assets	This provides details about the asset socio-economic status (SES) of households within the surveillance. The household SES dataset includes 117 asset indicators	117	705 984	N/A	N/A	251 629 (35.64)	N/A	N/A	454 355 (64.36)

(continued)

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

Table 2 (continued)

Dataset	Description	Variable count	Number of records (*N*)	Agincourt HDSS, South Africa			NUHDSS, Nairobi		
				Male (%)	Female (%)	*N* (%)	Male (%)	Female (%)	*N* (%)
Verbal autopsy (VA)	The VA dataset includes data from the WHO standard questionnaire for all deaths occurring among the surveillance population, including InsilicoVA causes of death assignment	367	26 743	10 904 (51.26)	10 361 (48.71)	21 271 (79.54)	2329 (42.56)	3137 (57.33)	5472 (20.46)
NCD indicators	The NCD indicators dataset consists of observations from harmonized datasets from the 10 independent health-related longitudinal studies nested within the surveillance population. The data specification identified 396 variables across the different studies. Table 1 presents a detailed description of each of the studies integrated as part of this dataset	396	130 550	42 248 (39.64)	64 337 (60.36)	106 585 (81.64)	11 688 (48.77)	12 277 (51.23)	23 965 (18.36)

For the 'Individuals' and 'Verbal autopsy (VA)' datasets, 'Number of records (*N*)' refers to unique individuals, whereas it refers to unique household records for the 'Households' dataset. The 'Number of records (*N*)', however, refers to observations captured at each visit for the 'Individual events', 'Individual socio-economic (SES) indicators', 'Household memberships', 'Household assets', and 'NCD indicators' datasets.
N/A, not applicable; WHO, World Health Organization.

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

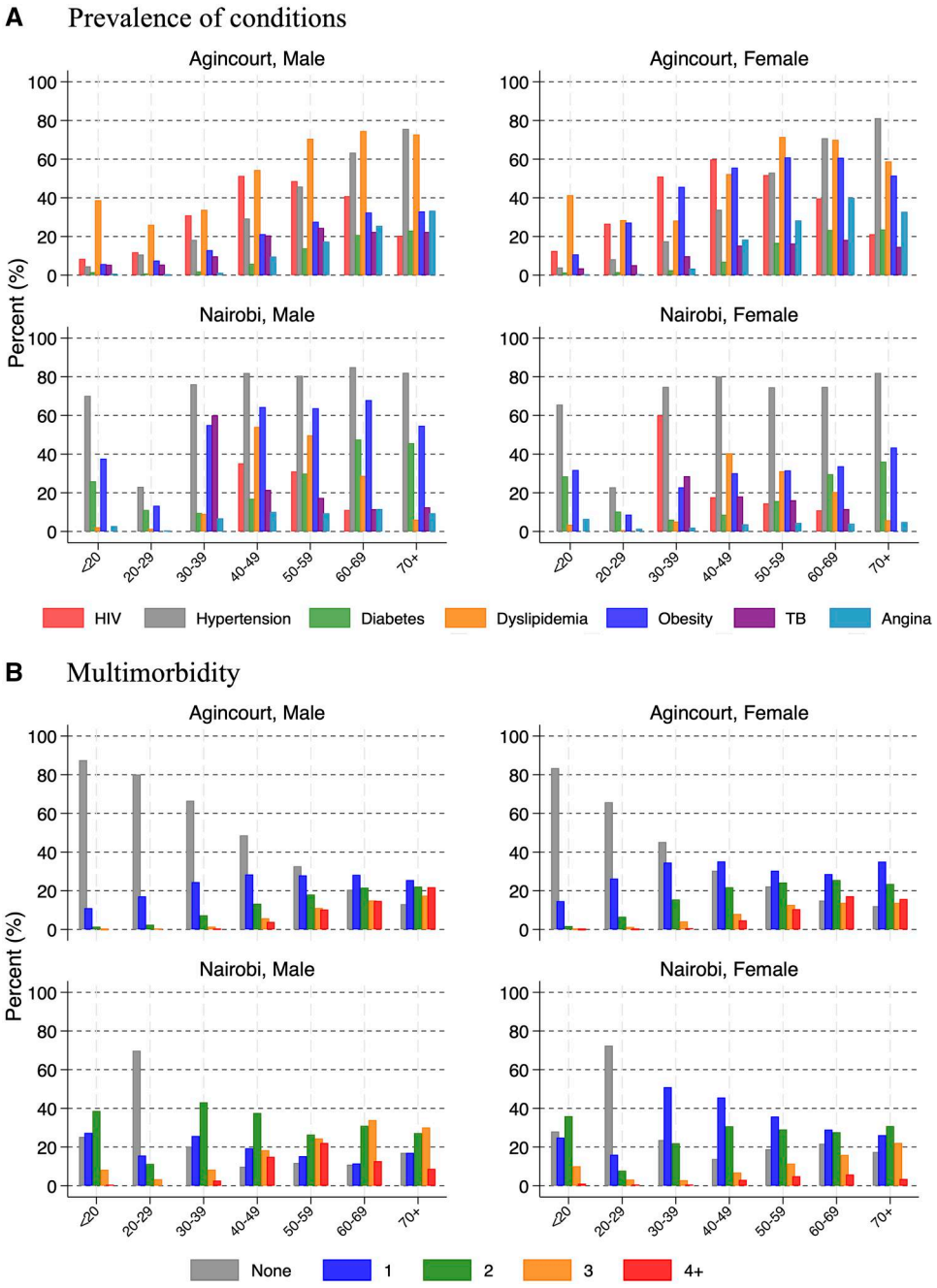


Figure 2 Age- and sex-specific distribution of chronic conditions and multimorbidity across sites. (a) Prevalence of selected chronic conditions (HIV, hypertension, diabetes, dyslipidaemia, obesity, tuberculosis, and angina) by age group and sex in Agincourt and Nairobi Urban surveillance sites. (b) Distribution of the number of chronic conditions per individual by age group and sex in Agincourt and Nairobi Urban surveillance sites.

Ongoing and planned analyses also include the application of machine-learning approaches, including auto-stratification techniques, to identify subpopulations with distinct risk profiles and disease trajectories. These methods are being used to explore the clustering of risk factors and conditions, improve the prediction of NCD outcomes, and identify groups at higher risk of adverse health outcomes. Further planned work includes analyses of mortality patterns, as well as disease-specific investigations aimed at improving the prediction and characterization of NCD risk and multimorbidity.

Beyond these, the harmonized dataset provides a valuable platform for addressing broader questions in epidemiology, population health, and health systems research. By integrating data from multiple longitudinal studies nested within HDSS platforms, the resource enables comparative and longitudinal analyses that are difficult to undertake by using individual studies alone.

Chronic conditions and multimorbidity

Among the 67 478 unique individuals who contributed to the 130 550 NCD indicator dataset observations, 53% (35 876) suffered from at least one of the following conditions: HIV (19.7%),

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

hypertension (35.6%), diabetes (6.3%), heart disease/failure (3.0%), stroke (6.0%), dyslipidaemia (43.9%), TB (3.4%), any cancers (4.8%), angina (12.8%), kidney disease/failure (7.3%), liver disease (1.2%), asthma (11.5%), thyroid disease (2.4%), and obesity (34.1%). Figure 2a highlights the prevalence of the top seven conditions and their differences by sex, age, and the respective HDSS sites. The figure generally shows increasing prevalence with age, particularly from the Agincourt HDSS. Furthermore, as many as 27% have two or more conditions, with observed differences across sex and age groups. Figure 2b summarizes the differences by age and sex, and Supplementary Table 3 provides details on the number of conditions across various sociodemographic characteristics. Supplementary Figure 3 presents the counts for the top 25 disease combinations.

Strengths and weaknesses

Strengths

The harmonized MADIVA database represents one of the first coordinated efforts to integrate longitudinal health and sociodemographic data across multiple population-based studies in South Africa and Kenya. The inclusion of both rural and peri-urban populations allows contextual comparisons and insights into geographic disparities in multimorbidity, including variation across demographic and socio-economic groups within the two countries. The standardization of core variables across distinct and independent studies enhances the analytic consistency of the dataset and supports cross-site, cross-country, and longitudinal analyses. The dataset is well suited for tracking multimorbidity patterns over time, investigating disease clustering, and informing context-specific public health interventions. An additional strength of the MADIVA database is the inclusion of biomarker and biological measurement data across contributing studies. These include blood pressure, blood glucose, lipid profiles, and anthropometric indicators—core variables for defining cardiovascular and metabolic risk factors central to the study of multimorbidity. This makes the MADIVA database a valuable population-based resource for epidemiological analyses, including disease clustering and risk stratification. Moreover, the integration of the HDSS dataset adds depth to the database by incorporating longitudinal socio-economic indicators at both individual and household levels. This facilitates nuanced exploration of the interplay between socio-economic status and the burden, distribution, and care pathways associated with multimorbidity in diverse African contexts. In addition, the dataset can reliably be used for mortality analysis across both sites (2002–15 for Nairobi Urban HDSS and 1992 to date for the Agincourt HDSS). Furthermore, work is ongoing to map the database to the Observational Medical Outcomes Partnership Common Data Model.

Weaknesses

The limitations of the MADIVA database primarily stem from the inherent challenges of harmonizing heterogeneous measures across independent studies conducted in distinct geographic and cultural settings. Additionally, variation exists in the collection of certain biomarkers—such as those used to define diabetes—due to differences in the fasting-status protocols, measurement techniques (e.g. point-of-care versus laboratory-based testing), and the instruments used across the different studies and sites. These methodological differences may influence cross-site comparability. To address this, we have systematically documented all survey-specific measurement nuances in the data-specification document to support the transparent interpretation of findings.

Self-reported measures can be influenced by recall and reporting biases, which are affected by respondents' education or health literacy. Although this may have led to the underestimation of disease prevalence or care-seeking behaviours, particularly in resource-limited settings, most studies included follow-up questions on treatment and care, to enhance accuracy.

Data resource access

Access to the MADIVA data is available with approval from the Research Ethics Committee at the University of the Witwatersrand, South Africa. Researchers wishing to access this data resource may contact the principal investigator, Professor Scott Hazelhurst (Scott.Hazelhurst@wits.ac.za). The data dictionary for the integrated, harmonized, de-identified participant-level MADIVA database, including the database schema and data-specification document, is available at http://github.com/MADIVA-DSI/data. Additional information about MADIVA, including details on collaboration and data access, is also provided at https://www.madiva.africa/.

Ethics approval

The University of the Witwatersrand Human Research Ethics Committee (Medical) (HREC M210825) and the Kenyan AMREF Research and Ethics and Scientific Review Committee (AMREF-ESRC P1206/2022) have approved MADIVA's projects.

Acknowledgements

We would like to thank each of the country-level study teams and study participants who made the compilation of this dataset possible. Members of the MADIVA Research Hub and the DS-I Africa Consortium are Daniel Ohene-Kwofie, Nkosinathi Masilela, Kerry Glover, Samuel Iddi, Daniel Maina, Molulaqhooa Linda Maoyi, Theophilous Mathema, Mwiza Gideon Singini, Scott Hazelhurst, and Chodziwadziwa Whiteson Kabudula.

Author contributions

D.O. led the writing of this manuscript. M.G.S. contributed to the initial draft of the paper. K.G. coordinated the teams regarding the harmonization process. C.K. and S.I. developed the harmonization framework and provided critical feedback on the data-quality and validation procedures. M.L.M. contributed to data-quality management. N.K., D.M., and T.M. led the harmonization and integration of datasets across the studies. S.I. and M.G.S. conducted initial data summaries and visualization for the resource profile. S.H. and

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

C.K. supervised the overall project and provided critical inputs to the harmonization process. All authors participated in the data-harmonization process, contributed to manuscript revisions, and approved the final version.

Supplementary material

Supplementary material is available at *IJE* online.

Conflicts of interest

The authors have no conflict of interest to declare.

Funding

Research reported in this publication was supported by the Fogarty International Center and National Institute of Biomedical Imaging and Bioengineering (NIBIB), the OD/Office of Strategic Coordination (OSC), and the Office of Data Science Strategy (ODSS) of the National Institutes of Health (under award number U54 TW 012077) and a Strategic Health Innovation Partnerships grant from the South African Medical Research Council (grant number N/A). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Data availability

See 'Data resource access', above.

Use of artificial intelligence (AI) tools

The authors performed all analyses and manuscript preparation independently, without utilizing any AI tools for language enhancement, data collection or analysis, result generation, or image production.

References

1. Mudie K, Jin MM, Tan *et al.* Non-communicable diseases in sub-Saharan Africa: a scoping review of large cohort studies. *J Glob Health* 2019;**9**:20409. https://doi.org/10.7189/jogh.09.020409
2. Kamp M, Achilonu O, Kisiangani I *et al.* Multimorbidity in African ancestry populations: a scoping review. *BMJ Glob Heal* 2023;**8**:e013509. https://doi.org/10.1136/bmjgh-2023-013509
3. World Health Organization (WHO). *Multimorbidity: Technical Series on Safer Primary Care*. Geneva, Switzerland: World Health Organization, 2016.
4. Nunes BP, Flores TR, Mielke GI, Thumé E, Facchini LA. Multimorbidity and mortality in older adults: a systematic review and meta-analysis. *Arch Gerontol Geriatr* 2016;**67**:130–8. https://doi.org/10.1016/j.archger.2016.07.008
5. Fortin M, Lapointe L, Hudon C, Vanasse A, Ntetu AL, Maltais D. Multimorbidity and quality of life in primary care: a systematic review. *Health Qual Life Outcomes* 2004;**2**:51. https://doi.org/10.1186/1477-7525-2-51
6. Cheng GJ, Wagner AL, O'Shea BQ, Joseph CA, Finlay JM, Kobayashi LC. Multimorbidity and mental health trajectories among middle-aged and older U.S. adults during the COVID-19 pandemic: longitudinal findings from the COVID-19 coping study. *Innov Aging* 2022;**6**:igac047. https://doi.org/10.1093/geroni/igac047
7. Fortin M, Bravo G, Hudon C, Lapointe L, Dubois M-F, Almirall J. Psychological distress and multimorbidity in primary care. *Ann Fam Med* 2006;**4**:417 LP–22. https://doi.org/10.1370/afm.528
8. Calderón-Larrañaga A, Poblador-Plou B, González-Rubio F, Gimeno-Feliu LA, Abad-Díez JM, Prados-Torres A. Multimorbidity, polypharmacy, referrals, and adverse drug events: are we doing things well? *Br J Gen Pract* 2012;**62**:e821 LP–e826. https://doi.org/10.3399/bjgp12X659295
9. Wolff JL, Starfield B, Anderson G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Arch Intern Med* 2002;**162**:2269–76. https://doi.org/10.1001/archinte.162.20.2269
10. Xu X, Mishra GD, Jones M. Mapping the global research landscape and knowledge gaps on multimorbidity: a bibliometric study. *J Glob Health* 2017;**7**:10414. https://doi.org/10.7189/jogh.07.010414
11. Micklesfield LK, Munthali R, Agongo G *et al.* Identifying the prevalence and correlates of multimorbidity in middle-aged men and women: a cross-sectional population-based study in four African countries. *BMJ Open* 2023;**13**:e067788. https://doi.org/10.1136/bmjopen-2022-067788
12. Cavve BS. Difficulties in data harmonisation: An examination of Amos (2024). *Australas Psychiatry* 2025;**33**:575–7. https://doi.org/10.1177/10398562241290028
13. Adhikari K, Patten SB, Patel AB *et al.* Data harmonization and data pooling from cohort studies: a practical approach for data management. *Int J Popul Data Sci* 2021;**6**:1680.
14. Manne-Goehler J, Theilmann M, Flood D *et al.* Data Resource Profile: The Global Health and Population Project on Access to Care for Cardiometabolic Diseases (HPACC). *Int J Epidemiol* 2022;**51**:e337–e349. https://doi.org/10.1093/ije/dyac125
15. Carrel M, Rennie S. Demographic and health surveillance: longitudinal ethical considerations. *Bull World Health Organ* 2008;**86**:612–6. https://doi.org/10.2471/blt.08.051037
16. Kahn K, Collinson MA, Xavier Gómez-Olivé F *et al.* Profile: Agincourt health and socio-demographic surveillance system. *Int J Epidemiol* 2012;**41**:988–1001. https://doi.org/10.1093/ije/dys115
17. Beguy D, Elung'ata P, Mberu B *et al.* Health and demographic surveillance system profile: The Nairobi Urban Health and Demographic Surveillance System (NUHDSS). *Int J Epidemiol* 2015;**44**:462–71. https://doi.org/10.1093/ije/dyu251
18. DDI Alliance. Data Documentation Initiative (DDI) Codebook: Version 2.5. *Iowa City, IA DDI Alliance* 2021. Available from: https://ddialliance.org/ (9 November 2025, date last accessed).
19. Norwegian Social Science Data Services (NSD). Nesstar Publisher, Version 4.0 *Bergen, Norw. NSD - Nor. Cent. Res. Data* 2008. Available from: https://www.ihsn.org/software/ddi-metadata-editor (9 November 2025, date last accessed).

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026

20. Adetunji KE, Mathema T, Kisiangani I *et al.* A Stratification Method for Identifying Subgroups at High-Risk for Type 2 Diabetes in sub-Saharan Africa. *Nat Commun* 2026;May 19; https://doi.org/10.1038/s41467-026-73226-6
21. Holden CC, Mdewa W, Mathema T *et al.* The economic cost of outpatient primary care of adults with multimorbidity (HIV, diabetes, and hypertension) in rural South Africa. *Health Policy Plan. Oxford Academic* 2026;**41**:570–83. https://doi.org/10.1093/heapol/czag016
22. Oshingbesan A, Kamp M, Mpangase PT *et al.* Sub-population identification of multimorbidity in sub-Saharan African populations. *Sci Rep* 2025;**15**:13992. https://doi.org/10.1038/s41598-025-96569-4
23. Wamukoya M, Kadengye DT, Iddi S, Chikozho C. The Nairobi Urban Health and Demographic Surveillance of slum dwellers, 2002–2019: value, processes, and challenges. *Glob Epidemiol* 2020;**2**:100024.https://doi.org/10.1016/j.gloepi.2020.100024
24. Byass P, Hussain-Alkhateeb L, D'Ambruoso L *et al.* An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. *BMC Med. BioMed Central Ltd* 2019;**17**:1–12. https://doi.org/10.1186/s12916-019-1333-6
25. Tluway F, Agongo G, Baloyi V *et al.* Cohort Profile: Africa Wits-INDEPTH partnership for Genomic studies (AWI-Gen) in four sub-Saharan African countries. *Int J Epidemiol* 2025; **54**:dyae173.
26. Ramsay M, Crowther N, Tambo E *et al.* H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob Heal Epidemiol Genomics*. 2016;**1**:e20.
27. Fabian J, Gondwe M, Mayindi N *et al.* Chronic kidney disease (CKD) and associated risk in rural South Africa: a population-based cohort study. *Wellcome Open Res* 2022;**7**:236. https://doi.org/10.12688/wellcomeopenres.18016.2
28. Xavier Gómez-Olivé F, Montana L, Wagner RG *et al.* Cohort profile: Health and ageing in Africa: A longitudinal study of an indepth community in South Africa (HAALSI). *Int J Epidemiol* 2018;**47**:689–690j. https://doi.org/10.1093/ije/dyx247
29. Gómez-Olivé FX, Nicole A, Brian H *et al.* Prevalence of HIV among those 15 and older in rural South Africa. *AIDS Care* 2013;**25**:1122–8. https://doi.org/10.1080/09540121.2012.750710
30. Thorogood M, Goudge J, Bertram M *et al.* The Nkateko health service trial to improve hypertension management in rural South Africa: Study protocol for a randomised controlled trial. *Trials* 2014;**15**:435. https://doi.org/10.1186/1745-6215-15-435
31. Haregu TN, Wekesah FM, Oti S, Egondi T, Kyobutungi C. The connection between non-communicable disease risk factors and risk perception among urban slum dwellers in Nairobi, Kenya. *African Popul Stud* 2016;**30**:17–23.
32. Oti SO, Vijver S V D, Kyobutungi C *et al.* A community-based intervention for primary prevention of cardiovascular diseases in the slums of Nairobi: the SCALE UP study protocol for a prospective quasi-experimental community-based trial. *Trials* 2013;**14**:409. https://doi.org/10.1186/1745-6215-14-409
33. Sanya RE, Karugu CH, Iddi S *et al.* Feasibility and impact of a patient support group care model on diabetes and hypertension care in informal settlements in Nairobi, Kenya: a quasi-experimental study. *Glob Health Action* 2025;**18**:2482304. https://doi.org/10.1080/16549716.2025.2482304

Downloaded from academic.oup.com/ije/article/55/4/dyag124/8741957 by University of Ghana user on 01 September 2026