International Journal of Population Data Science

Journal Website: www.ijpds.org

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Data Resource Profile: Harmonisation of a multimodal dataset to evaluate adolescent mental health in rural South Africa

Nondumiso Mthiyane[1,2], Edwin Mkwanazi[1], Patrick N Mwangala[3], Dickman Gareta[1,4,5], Sweetness Dube[1], Kobus Herbst[1,6], Maryam Shahmanesh[1,2], Kathy Baisley[1,7], and Amina Abubakar[3]

Submission History

Submitted:	11/05/2025
Accepted:	16/09/2025
Published:	08/12/2025

[1] Africa Health Research Institute, via R618 to Hlabisa, Somkhele, Mtubatuba, South Africa

[2] Institute for Global Health, University College London, Gower Street, London, WC1E 6BT, United Kingdom

[3] Institute for Human Development, Aga Khan University, 1st Parklands Avenue, Park Place Building, Nairobi, Kenya

[4] Institute of Social and Preventive Medicine, University of Bern, Mittelstrasse 43, Bern, Switzerland

[5] Graduate School for Health Sciences, University of Bern, Mittelstrasse 43, Bern, Switzerland

[6] DSTI-SAMRC South African Population Research Infrastructure (SAPRIN), South African Medical Research Council, 491 Peter Mokaba Road, Overport, Durban, South Africa

[7] Faculty of Epidemiology and Population Health, London School of Tropical Hygiene and Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

Abstract

Mental health disorders among adolescents and young adults in Africa are a growing concern, with most cases remaining undiagnosed or untreated due to limited resources. As the youth population increases, mental health issues are expected to rise, emphasising the need for targeted interventions. However, the lack of longitudinal data hinders researchers from understanding how social, behavioural, and clinical factors interact, which is essential for developing effective interventions. This study aimed to create a mental health data resource accessible to mental health researchers across Africa.

We identified five HIV prevention studies conducted within the Africa Health Research Institute's Health and Demographic surveillance area that collected mental health data from adolescents and young adults. These were combined with data from annual household surveys and routine clinic data. We extracted mental health data and relevant exposure variables (socio-demographics, behavior, general health, and clinic attendance) and harmonised them into a common format. The harmonisation process was conducted in accordance with the six steps framework outlined by the Maelstrom Research guidelines. We performed the harmonisation using Julia, with data stored in Microsoft SQL Server. Two mental health screening tools were used across the studies: the 14-item Shona Symptoms Questionnaire and the Patient Health Questionnaire (PHQ-9). From these tools, we generated key outcomes, including probable common mental disorders, depressive symptoms, and suicidal ideation.

The harmonised dataset includes 6,253 participants aged 13–24 years, with up to six data points collected between 2012 and 2022. Approximately 43.8% are males, and 63% are aged 13–19 years. About 12.9% reported having experienced suicidal ideation, 23.9% and 4.6% were screened positive using SSQ-14 and PHQ-9, respectively. While this dataset provides rich longitudinal data, it has some limitations. These include differences in recall periods of the mental health screening tools, potential self-reporting bias due to stigma, variation in measurement depth across studies, and limited cultural adaptation of the tools. As such, users should interpret and apply the data with appropriate caution. Despite these limitations, it enables the investigation of mental health trajectories and associated risk factors over time, providing critical insights into how social and biological influences shape mental well-being as young people transition into adulthood.

Keywords

adolescent mental health; longitudinal data; South Africa; data resource

Key features

- The AHRI adolescent mental health data is a harmonised dataset derived from the Africa Health Research Institute (AHRI) population cohort studies. It was produced in collaboration with the UtiliZing health Information for Meaningful impact in East Africa through the Data Science (UZIMA-DS) team, and can be used to help understand some of the drivers of mental health in adolescents and young adults.
- Mental health disorders are common in adolescents and have detrimental effects on individual performance and physical health, and more than half of all mental health disorders in adulthood originate from childhood and adolescence. In Africa, there is paucity of data on adolescent mental health to inform the design and implementation of contextual interventions for adolescents.
- The dataset consists of data from 6,253 adolescents and young adults aged between 13 and 24 years from rural South Africa who were identified from five studies conducted between 2012 and 2022 that measured mental health disorders.
- Data available include depressive symptoms, probable common mental disorders and suicidal ideation, socio-demographics, violence experience, sexual behaviour, sexual and reproductive health and clinic visit data.
- The data can be merged using unique individual identifiers with other datasets collected within AHRI to increase the scope of exposure variables and will be expanded in future with data from new mental health studies.
- Data can be accessed through the AHRI data repository under the terms of the AHRI Data Access Policy https://data.ahri.org/index.php/catalog?page=1&sk=uzima&sort_by=rank&sort_order=desc&ps=15.

*Corresponding Author:
Email Address: Nondumiso.Mthiyane@ahri.org (Nondumiso Mthiyane)

https://doi.org/10.23889/ijpds.v8i6.3003

December 9, 2025 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Background

Mental health disorders significantly contribute to illness, disability, and mortality, and pose substantial risks for suicide [1–3]. In Africa, there has been an increase in years lost to disability due to mental health disorders [4, 5], a trend largely driven by socio-economic stressors and adverse social determinants of health such as poverty, inequality, and inadequate healthcare services [6–8]. The majority of common mental health disorders, such as depression, anxiety, and behavioural disorders, often emerge in adolescence and can predict long-term mental and physical health challenges [9]. With a burgeoning youth population competing for limited socio-economic opportunities, the prevalence of mental health issues in Africa is expected to escalate, driven in part by a shortage of mental health resources to diagnose and treat mental health issues [10].

Young people in Africa face a dual burden of HIV and mental health issues. Evidence from sub-Saharan African countries highlights a bidirectional relationship between HIV and mental health, influenced by social and health challenges including poverty, various forms of violence and reproductive health outcomes [11–14]. These factors heighten the vulnerability of young people to both HIV infection and mental health problems, potentially affecting their academic or work performance [15, 16]. The situation in sub-Saharan Africa is further complicated by inadequate mental health resources, including data, and the need for locally validated, culturally relevant screening tools to identify mental health problems. Furthermore, there is a shortage of longitudinal studies to better understand how social and biological factors affect adolescent mental health in African contexts. Most research is conducted in high-income countries, making it difficult to design effective strategies and interventions that are tailored to the unique challenges of sub-Saharan Africa. Addressing this research gap is crucial for developing context-specific strategies.

Several HIV prevention research studies conducted within the Africa Health Research Institute (AHRI) Health and Demographic Surveillance System (HDSS) have collected data on adolescent mental health, creating an opportunity to combine these datasets into a comprehensive mental health data resource. In this study, we harmonised data from five cohort studies that collected mental health outcomes (symptoms of common mental disorders, depression and suicidal thoughts) alongside a range of risk factors. Guided by a conceptual framework, we hypothesised that demographic characteristics such as age, sex and urbanicity and broader social factors (including education, food insecurity, experiences of violence, social support and migration) influence mental health outcomes directly and indirectly as illustrated in Figure 1. These social factors may also shape individual behaviour and health (including HIV risks and access to care), thereby increasing vulnerability to poor mental health. This pilot project sought to bring together various data sources into a single, unified resource, designed in accordance with Findable, Accessible, Interoperable and Reusable (FAIR) principles to ensure it is accessible to mental health researchers across Africa [17]. This data harmonisation project was conducted by AHRI in collaboration with the UtiliZing Health Information for Meaningful Impact in East Africa through Data Science (UZIMA-DS) researchers from the Aga Khan University, Kenya.

Methods

Study design

This data harmonisation project used existing data from five studies conducted within AHRI HDSS to create a harmonised mental health data resource.

Setting

The studies contained in the harmonised data resource were conducted in the AHRI HDSS in uMkhanyakude district, KwaZulu-Natal, South Africa. Since 2000, AHRI HDSS has covered a population of approximately 140,000 household members in over 20,000 households residing in an area of 845 km^2 [18]. The area is predominantly rural with one town with an approximate population of 30,000 people. AHRI conducts annual household-based surveys to collect information on births, deaths, and migration patterns among all household members. In addition, residents aged $\geq$15 years are invited to participate in an annual HIV serosurvey, and to complete a questionnaire on general health and sexual behaviour. These data are routinely linked to HIV treatment for individuals accessing HIV treatment and care and hospital admission, with all admissions coded and classified using the International Classification of Diseases, 10$^{\text{th}}$ Revision (ICD-10) for those admitted at the district hospital.

Data sources

We used data from five studies involving cohorts of adolescents and young adults: 1) the DREAMS (Determined Resilient, Empowered, AIDS-free, Motivated and Safe) study; 2) Multilevel; 3) Isisekelo Sempilo; 4) Thetha Nami and 5) HIV Treatment as Prevention (TasP) (Table 1). These studies followed participants for a range of exposures and health outcomes including mental health outcomes.

The DREAMS cohort study evaluated the impact of a combination of HIV prevention interventions on new HIV infections among adolescent girls and young women (AGYW) [19]. The Multilevel cohort study investigated factors that influence AGYW and male partners' uptake and adherence to HIV prevention strategies. Both studies used the same data collection tools, including the 14-item Shona Symptoms Questionnaire (SSQ-14) to measure probable common mental disorders (CMD). For the harmonised data resource, we considered baseline data and follow-up data collected at up to three time points.

The Isisekelo Sempilo study was a 2×2 factorial trial that evaluated a sexual and reproductive health (SRH) intervention among adolescents and adults aged 16–29 years [20]. Mental health outcomes were measured using the SSQ-14 and Patient Health Questionnaire-9 (PHQ-9), a 9-item screening tool used to assess and measure the severity of depressive symptoms. Since the trial was still ongoing during data harmonisation, only baseline data were included in the harmonised dataset.

Thetha Nami was a stepped-wedge cluster randomised controlled trial that examined social mobilisation by peer

Figure 1: Conceptual framework of factors influencing adolescent mental health

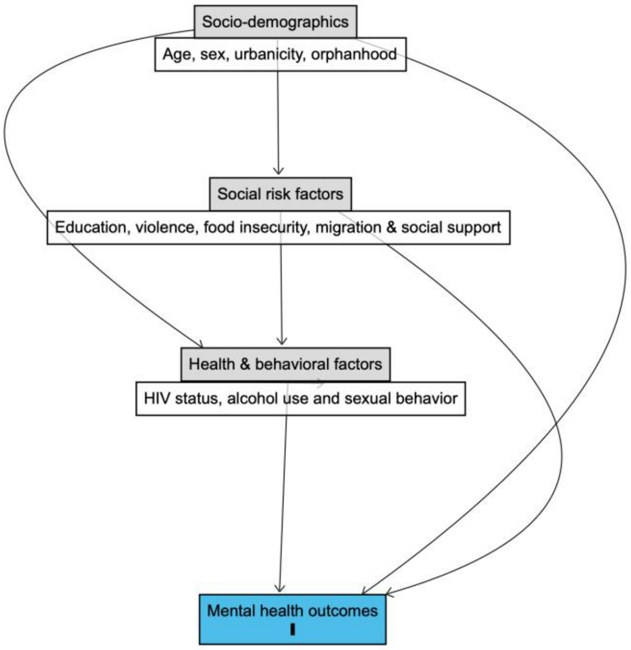

Socio-demographics
Age, sex, urbanicity, orphanhood
Social risk factors
Education, violence, food insecurity, migration & social support
Health & behavioral factors
HIV status, alcohol use and sexual behavior
Mental health outcomes

Table 1: Data sources and key variables

Data source	Study design	Participants	Period	Key Variables
DREAMS	Observational cohort study	Adolescent girls and young women aged 13-22	2017-19	Socio-demographics, food insecurity, violence, alcohol use, sexual behaviours, sexual and reproductive health, HIV status, mental health (probable common mental disorder)
Multilevel	Observational cohort study	Adolescent boys and young men aged 13-35, young women aged 25-29	2018-19	Socio-demographics, food insecurity, violence, alcohol use, sexual behaviours, sexual and reproductive health, HIV status, mental health (probable common mental disorder)
Isisekelo Sempilo	2 × 2 factorial randomised controlled trial	Adolescents and young adults aged 16-29	2020	Socio-demographics, sexually transmitted infections, HIV status, mental health (probable common mental disorder and depression)
Thetha Nami Ngithethe Nawe (Let's talk)	Stepped wedge cluster randomised trial	Adolescents and young adults aged 15-30	2022-ongoing	Socio-demographics, food insecurity, violence, alcohol use, sexual behaviours, sexual and reproductive health, HIV status, mental health (depression)
Treatment as Prevention (TasP)	Cluster randomised trial	Adolescents and young adults aged 15 or older	2012-2015	Socio-demographics, food insecurity, violence, alcohol use, sexual behaviours, HIV status, mental health (depression)

navigators to promote community-based sexual health and HIV care among young people aged 15-30 years [21]. Mental health data were collected using the PHQ-9. Again, only baseline data were included in the harmonised dataset because the trial was still ongoing during data harmonisation.

TasP was a cluster-randomised trial of the effect of HIV universal test and treat among individuals aged 16 years or older. The trial included mental health outcomes, and used the PHQ-9 to measure depression.

Data from these five studies were linked with the annual HDSS individual and household surveys, hospital admissions at the district hospital, and clinic visits at the 11 clinics serving the HDSS population, as recorded in the AHRILink.

Sampling methods in the original studies

Participants in DREAMS, Multi-level, Isisekelo Sempilo and Thetha Nami were randomly selected from the HDSS, with sampling stratified by age group, sex, and geographic area. Probability proportional to size sampling (PPS) was used to provide a self-weighting sample. In the TasP trial, participants were not randomly selected but all eligible household members were invited to participate in the study.

Harmonised data resource population

We used an expanded definition of adolescence (10-24 years) based on the culturally and contextually influenced delays in assuming traditional adult roles, such as financial independence and marriage [22, 23]. In many resource-limited settings, socio-economic factors such as unemployment, limited access to education and extended dependence on parents/caregivers often delay the transition from adolescence to adulthood. This approach recognises that age alone may not accurately capture the adolescent experience in these contexts. For the harmonised data resource, we restricted the sample to adolescents and young adults aged 13-24 years who were residents of the AHRI HDSS and participated in at least one of the cohort studies described in the previous section. Furthermore, participants were included in the harmonised data resource if they had data on mental health outcomes.

Data harmonisation process

We performed data harmonisation in six steps as shown in Figure 2. The data harmonisation process was informed by Maelstrom Research guidelines for rigorous retrospective data harmonisation [24].

Defining the research objectives

The primary aim of the study was to harmonise adolescent mental health data across five population-based studies conducted in the AHRI HDSS.

Assembling pre-existing knowledge and select studies

Study protocols, data dictionaries and questionnaires were collected for five studies. This was to ensure a clear understanding of how each variable was originally defined and measured across the five studies and to accurately identify comparable variables. Permission to use the datasets was obtained from the Principal Investigators of the primary studies.

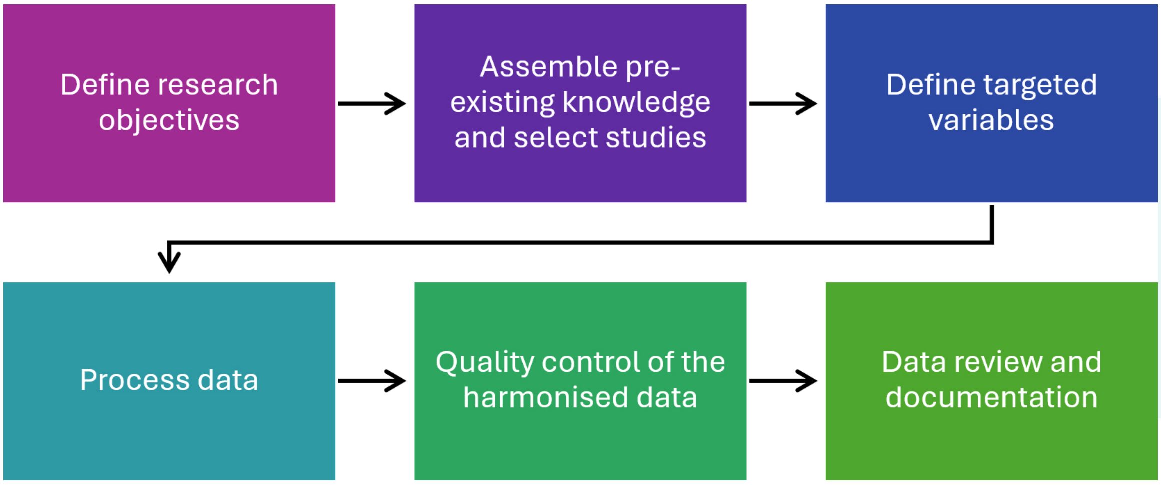
Defining targeted variables and evaluating harmonisation potential

The project team (Investigators from AHRI and AKU) held meetings to discuss data resources and identify mental health outcomes and exposures (potential risk factors). A list of socio-demographics, behaviour, general health and mental health variables, including the questions used to collect data was created (Supplementary Appendix 1). The study team reviewed the list to identify the screening tools used to collect mental health data and to decide which variables should be included in the harmonised dataset.

Prior to data harmonisation, the data were cleaned separately for each study. Outcomes and exposure variables were derived from raw data (where applicable) using the same format across studies. We considered three main mental health outcomes based on the available variables, two of which were generated based on the screening tools used. These included probable common mental disorder (CMD), depression, and suicidal ideation.

Probable CMD was assessed using the SSQ-14, a screening tool that was developed and validated in Zimbabwe [25, 26]. The SSQ-14 asks about symptoms experienced in the past seven days. While the tool has not yet been validated for use among adolescents under 18, it has been applied in sub-Saharan African contexts, including Tanzania [27], and has shown relatively low rates of misclassification compared to other screening tools such as the Self-Reporting Questionnaire (SRQ-20) [28, 29]. The outcome was measured among participants who responded to all SSQ-14 items (14 variables). All 'yes' responses were given a score of 1 and summed to create a final score for each individual. A binary variable indicating whether a participant's score was above a validated cut-off (>=9) was created. For participants who had missing items on the SSQ-14 (as described in Supplementary Appendix 2), the missing values were replaced with the mean score that was calculated among participants who completed the question. However, this was only done for participants who were missing not more than 20% of the questions (that is up to 3 questions out of 14); and those with missing data >20% were excluded. The 20% threshold for mean substitution on the SSQ-14 was based on commonly used practices in

Figure 2: Data harmonisation process


Define research objectives
Assemble pre-existing knowledge and select studies
Define targeted variables
Process data
Quality control of the harmonised data
Data review and documentation

handling missing data in psychometric scales. Previous studies have suggested that imputation methods, including mean substitution, may be acceptable when missingness is limited to 20% or less of items on a scale, as this level is unlikely to meaningfully alter scale reliability or validity [30, 31].

Depression was assessed using the PHQ-9 [32]. The PHQ-9 has been rigorously evaluated in multiple South African settings and evidence supports its reliability and validity in various South African contexts among adults and adolescents [33–35]. The PHQ-9 asks patients to rate, on a four-point scale ranging from “not at all” to “nearly every day,” the frequency with which they have experienced certain symptoms in the last two weeks. A total score was calculated by summing up the responses. For individuals with missing items, we replaced the missing values with the mean score calculated among participants with complete data. This was done for participants who were missing up to two questions out of nine [36]. A binary variable indicating whether an individual has depression was created, with a score of 10 or more indicated being depressed.

Suicidal ideation was measured using one of the following items from either the SSQ-14 or the PHQ-9:

- *In the past 7 days, did you at times feel like committing suicide?*

- *Thoughts that you would be better off dead or of hurting yourself in some way.*

Participants who gave an affirmative response to these questions were regarded as having suicidal ideation. For PHQ-9, “*several days*”,” *More than half the days*” or “*Nearly every day*” were all considered as an affirmative response. One study (Isisekelo Sempilo) used both SSQ-14 and PHQ-9 to measure mental health disorders; in this study, a participant was regarded as having suicidal ideation if they responded with “yes” to either of the tools.

Exposure variables (potential risk factors) included demographics (age, sex and household urbanicity), education, food insecurity, migration, social support, sexual behaviour (sexual and pregnancy history), experiences of violence, and HIV status. To ensure comparability across studies, we created binary composite variables for exposures. These variables were generated by combining multiple related items from different datasets into a single yes/no measure. For example, whether an individual was currently in school, had a history of food insecurity, had access to social support or met criteria for a particular experience or exposure. To manage heterogeneity, we applied clear criteria for inclusion in the harmonised variable:

- Variables had to align conceptually with the core definition of the construct.

- Minor differences in phrasing or time frame (e.g., ‘in the past year’ versus ‘ever’) were accepted as the underlying experience was meaningfully captured.

- In studies with multiple detailed items capturing different dimensions of the same construct (e.g., emotion, physical or sexual violence), a positive response to any of these items were coded as 1 in the binary composite variable.

- Studies that did not include any relevant item were excluded from the derivation of that specific composite variable.

For education, four studies had the same binary variable indicating whether a participant is currently enrolled in school. In TasP study, education status was generated from two other variables: currently employed and type of employment. Participants in TasP were coded as ‘currently in school’ if they reported being unemployed and studying, and as ‘not in school’ if they were neither studying nor employed.

For food insecurity, two variables from five studies were used to create a binary variable indicating whether a household had a history of food insecurity. Four of the studies used the same variable to measure food insecurity indicating whether any member of the household had ever skipped or cut the size of their meals due to financial reasons – while TasP study measured food insecurity based on whether a child in a household had missed a meal due to financial reasons.

Migration defined as whether a participant had ever migrated within or outside the HDSS was measured using surveillance data, so no standardisation was required. We considered migration episodes that occurred during adolescence (i.e., from the age of 10 years onward). In cases where both external and internal (within the HDSS) migrations occurred for the same individual, external migration was recorded as the final migration outcome. A composite migration variable with three categories (never migrated, internal migration (within HDSS), external migration) was generated.

For violence, a maximum of 15 detailed variables were used in four studies (DREAMS, Multilevel, Isisekelo Sempilo and Thetha Nami) to generate a binary variable indicating whether an individual had ever experienced any form of violence, including physical, emotional, or sexual violence. In TasP, three broad questions covering sexual and physical abuse were used to generate a binary variable.

For social support, 15 variables (11 variables from TasP and four variables from DREAMS, Multilevel, Isisekelo Sempilo and Thetha Nami) were used to create a binary variable indicating whether a participant had any form of social support including emotional, financial, informational and social interaction support.

Sexual behaviour was measured from variables indicating whether a participant has ever had sex or/and, for females, whether they had ever been pregnant. We created a composite variable with four categories namely; never had sex, ever had sex, ever been pregnant and unknown status (if participant had missing data or preferred not to answer a question).

Processing of data (harmonisation)

Following data preparation, the five study datasets were appended, and variables were aligned across data sources to ensure consistency and comparability. Outcome and exposure variables derived from different scoring scales were transformed to a common format, facilitating integration and analysis. (See Supplementary Appendix 1). For example, for suicidal ideation, we created variables indicating the source to specify whether the variable was calculated from the PHQ-9 or SSQ-14.

While both the SSQ-14 and PHQ-9 assess aspects of common mental disorders (CMD), they measure related but distinct constructs. The SSQ-14, developed specifically for sub-Saharan African populations, captures a broader spectrum of CMD symptoms, including anxiety and somatic symptoms. In contrast, the PHQ-9 focuses exclusively on depressive symptoms aligned with Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria. Due to these conceptual differences, we harmonised the data by treating the two instruments separately. This approach preserves the unique information provided by each tool and minimises the risk of misclassification that could occur if a combined score or crosswalk were created.

The entire harmonisation process was coded using Julia Programming Language including scoring methods and cutoffs [37, 38]. MSSQL Server Database was used for data storage during data transformation. The data preparation and harmonisation were performed by the second author. The link to the Julia code is publicly available on GitHub and can access via the following link: https://github.com/AHRIORG/MHDataHarmonization.

Quality control

To ensure reliability, quality control was performed independently by the first author who was not involved in the data harmonisation process. Quality control was performed using Stata 18 software, by summarising each variable with frequency tables in the harmonised data and comparing these with data summaries (frequencies) in the original datasets. All variables in the harmonised data were checked to ensure they are consistent with the data in the original studies. For example, if a study reported that 60% of participants were currently in enrolled in school, we cross-checked the harmonised variable to ensure it reflected a similar distribution according to any differences in coding or missing data. If there were any discrepancies, all datasets were reviewed in detail, and necessary corrections were made. These issues were discussed between the first and second author to understand how they arose and to agree on the most appropriate resolution. The same process was repeated until all issues were resolved.

Data review and documentation

After the data had undergone quality checks, all variables in the harmonised data were reviewed (i.e., checking variable and value labels, recoding variables from string to numeric) by the first and second authors. Data harmonisation documents (e.g., data dictionary) were reviewed and approved by the study team.

Results

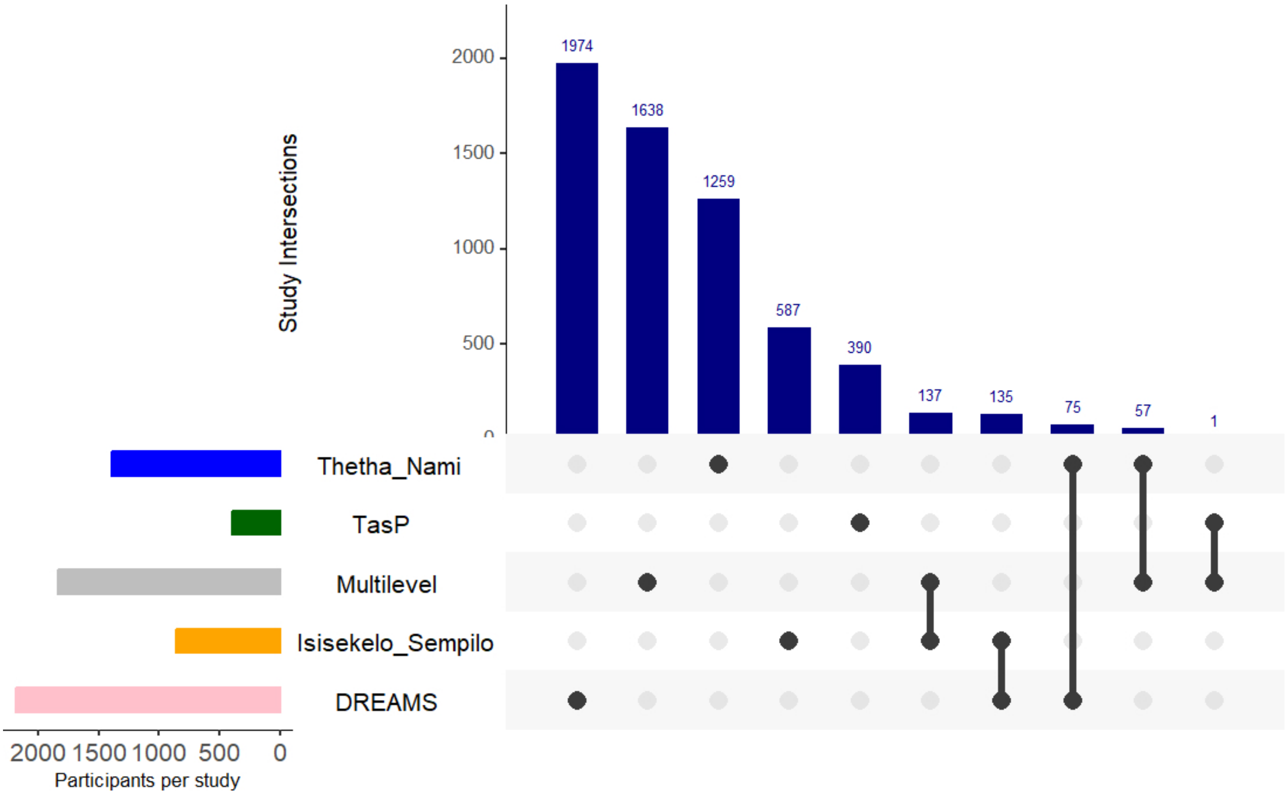
In Table 2, we provide a description of the number of individuals for whom harmonised data are available, along with the number of data points and the period for which variables are accessible. Data were available for most variables, with up to six data points. The majority (n=5,737) of participants had HIV test results and 19% had visited clinics within the HDSS.

A total number of 6,253 participants were included in the final harmonised dataset, of which 330 had participated in more than one study as indicated in Figure 3. DREAMS,

Table 2: Description of key variables

Variable	Description	Sample size	Period	Data points
Mental health outcomes				
suicidal_ideation	Thought of killing yourself	6051	2012-2022	6
depression_outcome	PHQ-9 score>=10	2131	2012-2014, 2021-2022	6
ssq14_outcome	SSQ-14 score>=9	4603	2017-2022	4
Exposures				
Migration	Ever migrated since age of 10	6253	2012-2022	1
Orphanstatus	Orphanhood status (one or both parents died)	6253	2012-2022	1
Education	Currently enrolled in school	6253	2012-2022	6
Governmenttgrant	Participant receives a social grant (child support, foster care)	6126	2012-2022	6
Foodsec	Skipped meal or reduced meal portions in the last 12 months	6253	2012-2022	6
Everhadsex	Ever had sex	6087	2012-2022	6
Violence	Ever experienced violence	6253	2012-2022	6
Everpregnant	Ever been pregnant	3172	2012-2022	4
Everdrankalcohol	Ever drank alcohol	6253	2012-2022	6
hiv_status	HIV test results	5737	2012-2022	6
hsv_test_result	Herpes Simplex virus type 2 test results	3863	2017-2022	4
Facilityservicename	Reasons for visiting a health facility	1190	2017-2022	+60
Condomlesssex	Ever had sex without a condom in the past 12 months	6253	2012-2022	6

Figure 3: The number of participants, by data source



Multilevel and Thetha Nami contributed the majority (64%) of the participants to the harmonised data.

Table 3 shows some of the characteristics of participants at baseline by data source. Of the 6,253 participants included in the dataset, 3,512 (56.2%) were female. About half of the participants had ever migrated, either within the HDSS or externally, since the age of 10. At the family level, more than a fifth of participants were orphaned (either single or double orphaned) and reported a history of food insecurity (i.e., ever having skipped or reduced meals in the past 12 months). More than a quarter had ever experienced violence and 8.9% tested positive for HIV.

Of 4,603 participants with SSQ-14 data, 1,101 (23.9%) had ever screened positive for probable CMD. Of 2,137 participants with PHQ-9 data, 4.6% had depressive symptoms. Of 6,051 participants, 12.9% reported having had suicidal ideation. Mental health disorders were more prevalent among females than males as shown (Figure 4).

Sensitivity analyses conducted using alternative SSQ-14 cut-off scores (≥8 and ≥7) among participants in the DREAMS and Multilevel studies (Supplementary Appendix 3), showed notable variation in the estimated prevalence of probable CMD, suggesting that prevalence may have been underestimated.

Discussion

Our study was motivated by the critical gap in longitudinal research exploring how social and biological factors influence adolescent mental health in African contexts. To address this, we developed a harmonised data resource that includes longitudinal mental health data for 6,253 adolescents and young people in rural South Africa. The harmonised dataset incorporates a wide range of relevant social and biological risk factors, such as food insecurity, social support, migration, orphanhood, exposure to violence, sexual behaviour and alcohol use. It also captures key health-related variables including HIV status and clinic attendance in facilities within the AHRI HDSS.

This harmonised dataset offers a valuable resource for researchers to examine adolescent mental health outcomes and exposures across multiple time points, enabling a more nuanced understanding of mental health trajectories and the factors affecting them. While only baseline data from studies such as Isisekelo Sempilo and Thetha Nami were available at the time of data harmonisation, it is important to note that participants from these studies may also be captured in other AHRI studies conducted subsequently, which could enable retrospective linkage and longitudinal analysis. By tracking changes (such as completing education, migrating, sexual behaviour) over time, researchers will be able to identify patterns, early warning signs, and key milestones in mental health development. This will not only enhance our understanding of how mental health evolves during adolescence but also provide insights into the impact of socio-economic, biological, and environmental factors on mental well-being. Furthermore, this dataset will serve as a foundation for new research initiatives, helping to identify effective intervention strategies, improve mental health policies, and guide the development of culturally relevant mental health programs tailored to the specific needs of adolescents in sub-Saharan Africa.

Table 3: Characteristics of participants at baseline, overall and by data source

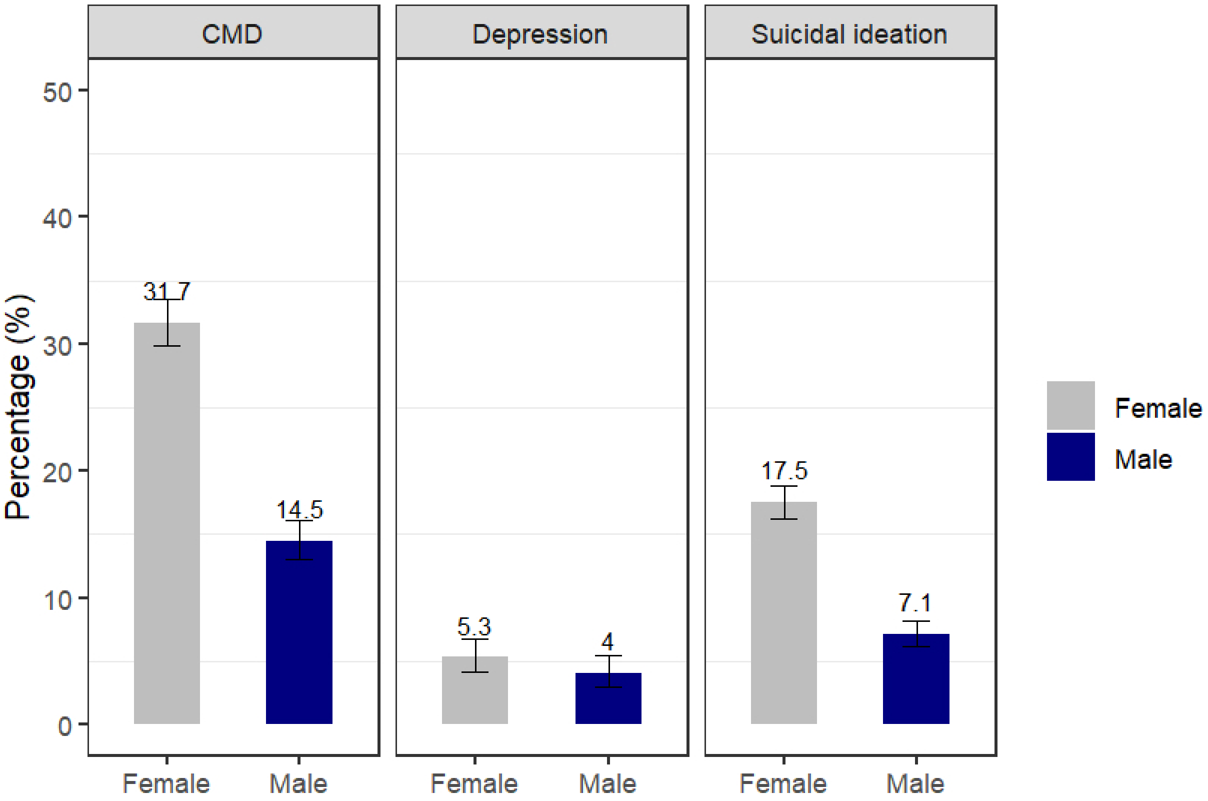
	All* N=6253	**DREAMS N=2184**	**Isisekelo Sempilo N=587**	**Multi-level N=1832**	**TasP N=391**	**Thetha Nami N=1259**
Socio-demographics						
Age, mean (SD)	18.4 (3.0)	17.2 (2.8)	18.0 (2.9)	18.9 (2.7)	20.8 (2.1)	21.3 (1.4)
Age group						
13-19	3939 (63.0)	1613 (73.9)	249 (42.4)	1265 (69.1)	94 (24.0)	718 (57.0)
20-24	2314 (37.0)	571 (26.1)	338 (57.6)	567 (30.9)	297 (76.0)	541 (43.0)
Participant's sex						
Male	2741 (43.8)		303 (51.6)	1776 (96.9)	54 (13.8)	608 (48.3)
Female	3512 (56.2)	2184 (100)	284 (48.4)	56 (3.1)	337 (86.2)	651 (51.7)
Ever migrated						
Never	3059 (49.3)	764 (35.0)	302 (51.5)	869 (47.5)	130 (36.8)	994 (79.3)
Within HDSS	621 (10.0)	246 (11.3)	64 (10.9)	201 (11.0)	32 (9.1)	78 (6.2)
External migration	2527 (40.7)	1173 (53.7)	220 (37.5)	761 (41.6)	191 (54.1)	182 (14.5)
One or both parents died						
Both Parents Alive	4925 (78.8)	1665 (76.2)	456 (77.7)	1406 (76.7)	268 (68.5)	1130 (89.8)
One Parent Deceased	1210 (19.4)	475 (21.7)	112 (19.1)	393 (21.5)	108 (27.6)	122 (9.7)
Both Parents Deceased	118 (1.9)	44 (2.0)	19 (3.2)	33 (1.8)	15 (3.8)	7 (0.6)
Skipped meal in the last 12 months						
No	4841 (77.4)	1545 (70.7)	433 (73.8)	1435 (78.3)	169 (43.2)	1259 (100.0)
Yes	1412 (22.6)	639 (29.3)	154 (26.2)	397 (21.7)	222 (56.8)	
Experienced violence						
No	4513 (72.2)	1425 (65.2)	475 (80.9)	1070 (58.4)	384 (98.2)	1159 (92.1)
Yes	1740 (27.8)	759 (34.8)	112 (19.1)	762 (41.6)	7 (1.8)	100 (7.9)
General health						
HIV test results						
Negative	2620 (41.9)	253 (11.6)	461 (78.5)	666 (36.4)	200 (51.2)	1040 (82.6)
Positive	559 (8.9)	218 (10.0)	62 (10.6)	90 (4.9)	92 (23.5)	97 (7.7)
Status Unknown	3074 (49.2)	1713 (78.4)	64 (10.9)	1076 (58.7)	99 (25.3)	122 (9.7)

*Individuals who participated in more than one study, the first baseline survey was considered.

Given the limitations of our data, users should apply the dataset with appropriate caution. First, none of the studies included young adolescents aged 10-12, however, the inclusion of young adults aged 20-24 allows for comparison as adolescents transition to adulthood. Second, the differences in the recall periods of the SSQ-14 (7 days) and PHQ-9 (14 days) could influence symptom reporting and affect comparability, both within our study and across other studies using these tools. The shorter 7-day recall period of the SSQ-14 may capture more transient or acute symptoms of psychological distress, while the PHQ-9's 14-day time frame reflects more persistent depressive symptoms. As such, participants may report different symptom patterns depending on the specific timeframe each tool targets. For instance, individuals experiencing recent stressors might be more likely to screen positive on the SSQ-14 but not meet the threshold for depression on the PHQ-9 if symptoms have not persisted over two weeks. These differences could lead to variability in prevalence estimates and in the degree of overlap between the two tools. When comparing results across studies, it is important to interpret findings with this nuance in mind. Third, we acknowledge the possibility of self-report bias, particularly due to stigma associated with mental health in some communities, which may have led to underreporting of symptoms. Fourth, in both studies with multiple detailed items and those with fewer broader questions, a positive response to any item was coded as '1'. e.g., for violence experiences. While this approach maximises comparability across studies, differences in item depth and sensitivity may affect measurement precision and should be considered when interpreting results. Lastly, although both PHQ-9 and SSQ-14 have been widely used in the Southern Africa, they may still not fully capture locally specific expressions of distress in this population and that future work should continue to explore more culturally grounded mental health screening tools.

The strength of this dataset is its flexibility, allowing for the addition of new variables and data as further research is conducted, ensuring it remains comprehensive and up to date. The data produced in this pilot study may be merged with the AHRI HDSS routine data and other nested studies, enabling the tracking of changes at both the individual and family levels and their impact on mental health outcomes. Additionally, participants in these cohort studies were exposed to various interventions at both the community and family

Figure 4: Prevalence of common mental disorders (with 95% confidence intervals), by sex



levels, including social protection and social asset building. The uptake of these interventions was also measured, presenting a valuable opportunity to expand the harmonised data by incorporating additional exposure variables linked to mental health.

Data access

Adolescent mental health data can be accessed through the AHRI data repository (https://data.ahri.org/index.php/catalog?page=1&sk=uzima&sort_by=rank&sort_order=desc&ps=15) after self-registration and completion of a short data-use agreement form available online. Data documentation, including technical documents and data dictionaries, is available in the repository.

Conclusions

The harmonised adolescent mental health dataset offers important insights into the mental health of adolescents and young adults in rural South Africa. This dataset serves as a vital resource for advancing research and intervention strategies and has the potential to influence mental health policies and programs tailored to the specific needs of adolescents in sub-Saharan Africa. It enables a more comprehensive understanding of the social, economic, and environmental factors impacting youth mental health. Furthermore, the development of this harmonised mental health dataset underscores the importance of culturally sensitive research approaches to improve mental well-being and resilience among young populations. Ensuring the use of culturally validated tools and locally informed methodologies is essential for generating reliable context-appropriate data that can support future mental health research and intervention development across diverse African settings.

Acknowledgements

The authors would like to acknowledge the AHRI research teams for their commitment to the studies as well as the UZIMA-DS team at the Aga Khan University team for their ongoing support on this pilot project. We also extend our appreciation to our research community for participating in our studies.

Ethics approval

This pilot project was approved by the University of KwaZulu-Natal (UKZN) Biomedical Research Committee (BREC). The reference number is BREC/00007187/2024. The original studies obtained ethical approval from the UKZN BREC. Potential participants were visited in their homes and invited to participate in the study. Participants aged 18 years or older provided written consent. Parental consent with participant written assent was sought for participants younger than 18 years. Participants in the primary studies also consented to have their data stored and used in future research.

Conflict of interest

The authors declare that they have no conflict of interest.

Publication consent

All participants in the original studies provided informed consent for the use and publication of their anonymised data. No identifying information is included in the dataset.

Funding

Research reported in this publication was supported by the Office of The Director, National Institutes of Health (OD), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the National Institute Of Mental Health (NIMH), and the Fogarty International Center (FIC) of the National Institutes of Health under award number U54TW012089 (Abubakar A and Waljee A). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Africa Health Research Institute is supported by core funding from the Wellcome Trust [grant number: 227167/A/23/Z]. This research was funded in whole, or in part, by the Wellcome Trust (grant numbers: 227167/A/23/Z and 210479/Z/18/Z). For the purpose of open access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript version arising from this submission. The funders have played no role in the study design, writing of the manuscript and in the decision to submit the manuscript for publication.

Data availability

The dataset described in this paper is available to qualified mental health researchers upon request through the AHRI Data Repository: (https://data.ahri.org/index.php/catalog?page=1&sk=uzima&sort_by=rank&sort_order=desc&ps=15).

References

1. WHO. Mental health of adolescents 2021 [Available from: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health.

2. Bachmann S. Epidemiology of Suicide and the Psychiatric Perspective. Int J Environ Res Public Health. 2018;15(7). https://doi.org/10.3390/ijerph15071425.

3. GBD 2017 Child and Adolescent Health Collaborators. Diseases, Injuries, and Risk Factors in Child and Adolescent Health, 1990 to 2017: Findings From the Global Burden of Diseases, Injuries, and Risk Factors 2017 Study. JAMA Pediatrics. 2019;173(6):e190337-e. https://doi.org/10.1001/jamapediatrics.2019.0337

4. Sankoh O, Sevalie S, Weston M. Mental health in Africa. The Lancet Global Health. 2018;6(9):e954-e5. https://doi.org/10.1016/S2214-109X(18)30303-6

5. WHO. Adolescent and young adult health 2021 [Available from: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions.

6. Lassi ZS, Salam RA, Bhutta ZA. Recommendations on Arresting Global Health Challenges Facing Adolescents and Young Adults. Annals of global health. 2017;83(5-6):704-12. https://doi.org/10.1016/j.aogh.2017.10.027i

7. Fox L, Senbet LW, Simbanegavi W. Youth Employment in Sub-Saharan Africa: Challenges, Constraints and Opportunities. Journal of African Economies. 2016;25(suppl_1):i3-i15. https://doi.org/10.1093/jae/ejv027

8. Remien RH, Stirratt MJ, Nguyen N, Robbins RN, Pala AN, Mellins CA. Mental health and HIV/AIDS: the need for an integrated response. AIDS. 2019;33(9):1411-20.

9. WHO. Adolescent mental health Geneva: WHO; 2020 [Available from: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health.

10. Nicholas A, Joshua O, Elizabeth O. Accessing Mental Health Services in Africa: Current state, efforts, challenges and recommendation. Ann Med Surg (Lond). 2022;81:104421. https://doi.org/10.1016/j.amsu.2022.104421

11. Bach JM, Louw D. Depression and exposure to violence among Venda and Northern Sotho adolescents in South Africa. African journal of psychiatry. 2010;13(1):25-35. https://doi.org/10.4314/ajpsy.v13i1.53426

12. Maharaj V, Tomita A, Thela L, Mhlongo M, Burns JK. Food Insecurity and Risk of Depression Among Refugees and Immigrants in South Africa. Journal of immigrant and minority health. 2017;19(3):631-7. https://doi.org/10.1007/s10903-016-0370-x

13. Osok J, Kigamwa P, Stoep AV, Huang KY, Kumar M. Depression and its psychosocial risk factors in pregnant Kenyan adolescents: a cross-sectional study in a community health Centre of Nairobi. 2018;18(1):136. https://doi.org/10.1186/s12888-018-1706-y

14. Stansfeld SA, Rothon C, Das-Munshi J, Mathews C, Adams A, Clark C, et al. Exposure to violence and mental health of adolescents: South African Health and Well-being Study. BJPsych open. 2017;3(5):257-64. https://doi.org/10.1192/bjpo.bp.117.004861

15. Das JP. Unity and diversity of views on intelligence and consciousness: Where the East meets the West. Psychological Studies. 2009;54(1):38-41. https://doi.org/10.1007/s12646-009-0005-6

16. Gupta J, Vegelin C. Sustainable development goals and inclusive development. International Environmental Agreements: Politics, Law and Economics. 2016;16(3):433-48. https://doi.org/10.1007/s10784-016-9323-z

17. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018. https://doi.org/10.1038/sdata.2016.18

18. Gareta D, Baisley K, Mngomezulu T, Smit T, Khoza T, Nxumalo S, et al. Cohort profile update: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. International Journal of Epidemiology. 2021. https://doi.org/10.1093/ije/dyaa264

19. Birdthistle I, Schaffnit SB, Kwaro D, Shahmanesh M, Ziraba A, Kabiru CW, et al. Evaluating the impact of the DREAMS partnership to reduce HIV incidence among adolescent girls and young women in four settings: a study protocol. BMC public health. 2018;18(1):912. https://doi.org/10.1186/s12889-018-5789-7

20. Chidumwa G, Chimbindi N, Herbst C, Okeselo N, Dreyer J, Zuma T, et al. Isisekelo Sempilo study protocol for the effectiveness of HIV prevention embedded in sexual health with or without peer navigator support (Thetha Nami) to reduce prevalence of transmissible HIV amongst adolescents and young adults in rural KwaZulu-Natal: a 2 × 2 factorial randomised controlled trial. BMC public health. 2022;22(1):454. https://doi.org/10.1186/s12889-022-12796-8

21. Busang J, Zuma T, Herbst C, Okesola N, Chimbindi N, Dreyer J, et al. Thetha Nami ngithethe nawe (Let's Talk): a stepped-wedge cluster randomised trial of social mobilisation by peer navigators into community-based sexual health and HIV care, including pre-exposure prophylaxis (PrEP), to reduce sexually transmissible HIV amongst young people in rural KwaZulu-Natal, South Africa. BMC public health. 2023;23(1):1553. https://doi.org/0.1186/s12889-023-16262-x

22. Kinghorn A, Shanaube K, Toska E, Cluver L, Bekker L-G. Defining adolescence: priorities from a global health perspective. The Lancet Child & Adolescent Health. 2018;2(5):e10. https://doi.org/10.1016/S2352-4642(18)30096-8

23. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. The Lancet Child & Adolescent Health. 2018;2(3):223-8. https://doi.org/10.1016/S2352-4642(18)30022-1

24. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2017;46(1):103-5. https://doi.org/10.1093/ije/dyw075

25. Chibanda D, Verhey R, Gibson LJ, Munetsi E, Machando D, Rusakaniko S, et al. Validation of screening tools for depression and anxiety disorders in a primary care population with high HIV prevalence in Zimbabwe. Journal of affective disorders. 2016;198:50-5. https://doi.org/10.1016/j.jad.2016.03.006

26. Patel V, Simunyu E, Gwanzura F, Lewis G, Mann A. The Shona Symptom Questionnaire: the development of an indigenous measure of common mental disorders in Harare. Acta psychiatrica Scandinavica. 1997;95(6):469-75. https://doi.org/10.1111/j.1600-0447.1997.tb10134.x

27. Uriyo JG, Abubakar A, Swai M, Msuya SE, Stray-Pedersen B. Prevalence and correlates of common mental disorders among mothers of young children in Kilimanjaro Region of Tanzania. PloS one. 2013;8(7):e69088. https://doi.org/10.1371/journal.pone.0069088

28. Haney E, Singh K, Nyamukapa C, Gregson S, Robertson L, Sherr L, et al. One size does not fit all: psychometric properties of the Shona Symptom Questionnaire (SSQ) among adolescents and young adults in Zimbabwe. Journal of affective disorders. 2014;167:358-67. https://doi.org/10.1016/j.jad.2014.05.041

29. Winston M, Smith J. A trans-cultural comparison of four psychiatric case-finding instruments in a Welsh community. Soc Psychiatry Psychiatr Epidemiol. 2000;35(12):569-75. https://doi.org/10.1007/s001270050281

30. Dong Y, Peng CY. Principled missing data methods for researchers. Springerplus. 2013;2(1):222. https://doi.org/10.1186/2193-1801-2-222

31. Siddiqui O. Methods for Computing Missing Item Response in Psychometric Scale Construction. American Journal of Biostatistics. 2015;5:1-6. https://doi.org/10.3844/amjbsp.2015.1.6

32. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-13. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

33. Bhana A, Rathod SD, Selohilwe O, Kathree T, Petersen I. The validity of the Patient Health Questionnaire for screening depression in chronic care patients in primary health care in South Africa. BMC Psychiatry. 2015;15:118. https://doi.org/10.1186/s12888-015-0503-0

34. Hart C, Draper CE, Soepnel LM, Godongwana M, Mabetha K, Nyati LH, et al. Examining the psychometric properties of the Patient Health Questionnaire-9 and Generalized Anxiety Disorder-7 among young urban South African women. J Affect Disord. 2025;369:61-70.https://doi.org/10.1016/j.jad.2024.09.145

35. Rakshasa-Loots AM, Hamana T, Fanqa B, Lindani F, van Wyhe K, Kruger S, et al. isiXhosa translation of the Patient Health Questionnaire (PHQ-9) shows satisfactory psychometric properties for the measurement of depressive symptoms [Stage 2]. Brain Neurosci Adv. 2023;7:23982128231194452. https://doi.org/10.1177/23982128231194452

36. Kroenke K, Spitzer RL, Williams JB, Löwe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry. 2010;32(4):345-59. https://doi.org/10.1016/j.genhosppsych.2010.03.006

37. Bezanson J, Karpinski S, Shah V, Edelman A. Julia: A Fast Dynamic Language for Technical Computing 2012. https://doi.org/10.48550/arXiv.1209.5145

38. Roesch E, Greener JG, MacLean AL, Nassar H, Rackauckas C, Holy TE, et al. Julia for biologists. Nat Methods. 2023;20(5):655-64. https://doi.org/10.1038/s41592-023-01832-z

Abbreviations

AGYW:	Adolescent girls and young women
AHRI:	Africa Health Research Institute
AKU:	Aga Khan University
CMD:	Common mental disorder
DREAMS:	Determined Resilient Empowered AIDS-free Motivated and Safe
DSM :	Diagnostic and Statistical Manual of Mental Disorders
FAIR:	Findable, Accessible, Interoperable and Reusable
HDSS:	Health and demographic surveillance system
HSV:	Herpes simplex virus
ICD:	International Classification of Diseases
MSSQL:	Microsoft SQL Server
PHQ:	Patient Health Questionnaire
PPS:	Probability proportional to size sampling
SRH:	Sexual and reproductive health
SSQ:	Shona Symptoms Questionnaire
TasP:	Treatment as prevention
UZIMA-DS:	UtiliZing health Information for Meaningful impact in East Africa through Data Science

Supplementary appendices

Supplementary Appendix 1. List of target variables (including the questions used to measure them) used to generate a harmonised dataset

Supplementary Appendix 2. Descriptive analysis of missingness by data source

Supplementary Appendix 3. Sensitivity analyses