

# **Hackathon: An AI-driven Tool for Record Linkage in HDSS Communities within the INSPIRE Network**

## **Study Documentation**

January 25, 2024

# Metadata Production

Identification	HACKATHON.2024.V3
----------------	-------------------

# Table of Contents

<a href="#">Overview.....</a>	<a href="#">4</a>
<a href="#">Scope &amp; Coverage.....</a>	<a href="#">4</a>
<a href="#">Files Description.....</a>	<a href="#">5</a>
<a href="#">synthetic_hdss_v3.....</a>	<a href="#">5</a>
<a href="#">synthetic_facility_v3.....</a>	<a href="#">5</a>
<a href="#">Variables List.....</a>	<a href="#">6</a>
<a href="#">synthetic_hdss_v3.....</a>	<a href="#">6</a>
<a href="#">synthetic_facility_v3.....</a>	<a href="#">6</a>
<a href="#">Variables Description.....</a>	<a href="#">7</a>
<a href="#">synthetic_hdss_v3.....</a>	<a href="#">8</a>
<a href="#">synthetic_facility_v3.....</a>	<a href="#">10</a>

## Hackathon: An AI-driven Tool for Record Linkage in HDSS Communities within the INSPIRE Network

Overview	
Identification	HACKATHON.2024.V3
Version	V3
<p><b>Abstract</b></p> <p>Background:</p> <p>The Implementation Network for Sharing Population Information from Research Entities (INSPIRE) collaborates with over 15 Health and Demographic Surveillance Sites (HDSS) in 7 African countries. INSPIRE network is keen to devise novel means of improving record linkage at community level. Record linkage tools have the ability to identify a person who lives in the HDSS catchment area and link them to the social services that they receive in the community. The social services include health and education services in HDSS catchment area. Effective record linkage to social services in HDSS communities has the following potential advantages;</p> <ol style="list-style-type: none"> <li>1. Ability to correlate household social characteristics with health seeking behavior</li> <li>2. Can identify early solutions to improve quality of life at household and community level</li> <li>3. Support the prediction of early warning epidemics at community level</li> <li>4. Correlate household demographic social characteristics with school attendance among children living in catchment area</li> <li>5. Support economic modelling and further estimation of the SDG indicators at community level</li> </ol> <p>Proposed solution:</p> <p>The INSPIRE network leverages experience from collaborating institutions to launch a competitive hackathon to design an innovative approach to record linkage in the HDSS. The team aims to explore the use of artificial intelligence tools and advances in machine learning algorithms to design a record linkage application or tool. We hypothesize the use of a tool that can identify a member of the HDSS when they seek health services with the community.</p> <p>Requirements for linkage:</p> <p>For development of a novel and the effective record linkage tool (MIRAGE), key variables that include participant identifying information (PII) will be required. The table shows key variables requested from the HDSS databases</p>	

Scope & Coverage	
Countries	
Geographic Coverage	

# Files Description

Dataset contains 2 file(s)

synthetic_hdss_v3	
# Cases	4115
# Variable(s)	9
<b><u>File Content</u></b> Key variables from HDSS.  MD5 File Checksum: f5b1c29b95b42e44cece831a997c08d6	
<b><u>Version</u></b> V3	

synthetic_facility_v3	
# Cases	2902
# Variable(s)	9
<b><u>File Content</u></b> Key variables from health facility in HDSS catchment area.  MD5 File Checksum: 3e3b24508c78c54bd19b11b35de6df9c	
<b><u>Version</u></b> V3	

# Variables List

Dataset contains 18 variable(s)

File synthetic_hdss_v3							
#	Name	Label	Type	Format	Valid	Invalid	Question
1	<a href="#">recnr</a>	recnr	continuous	numeric.0	4115	0	-
2	<a href="#">firstname</a>	firstname	discrete	character-13	4115	0	-
3	<a href="#">lastname</a>	lastname	discrete	character-13	-	-	-
4	<a href="#">petname</a>	petname	discrete	character-12	-	-	-
5	<a href="#">dob</a>	dob	discrete	character	0	-	-
6	<a href="#">sex</a>	sex	discrete	numeric.0	4115	0	-
7	<a href="#">nationalid</a>	nationalid	discrete	character-4	4115	0	-
8	<a href="#">hdssid</a>	hdssid	discrete	character-6	4115	0	-
9	<a href="#">hdsshhid</a>	hdsshhid	discrete	character-8	4115	0	-

File synthetic_facility_v3							
#	Name	Label	Type	Format	Valid	Invalid	Question
1	<a href="#">recnr</a>	recnr	continuous	numeric.0	2902	0	-
2	<a href="#">firstname</a>	firstname	discrete	character-13	2902	0	-
3	<a href="#">lastname</a>	lastname	discrete	character-13	2902	0	-
4	<a href="#">petname</a>	petname	discrete	character-13	2902	0	-
5	<a href="#">dob</a>	dob	discrete	character	0	-	-
6	<a href="#">sex</a>	sex	discrete	numeric.0	2902	0	-
7	<a href="#">nationalid</a>	nationalid	discrete	character-10	2902	0	-
8	<a href="#">patientid</a>	patientid	continuous	numeric.0	2902	0	-
9	<a href="#">visitdate</a>	visitdate	discrete	character	0	-	-

# Variables Description

**Dataset contains 18 variable(s)**

File : synthetic_hdss_v3			
# recnr: recnr			
Information	[Type= continuous] [Format=numeric] [Range= 1-4115] [Missing=*]		
Statistics [NW/ W]	[Valid=4115 /-] [Invalid=0 /-]		
Definition	A sequential number to identify every record in the dataset uniquely		
# firstname: firstname			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=4115 /-] [Invalid=0 /-]		
Definition	First name of the Person as it appears in DHIS-2 (or equivalent of national health information system)		
# lastname: lastname			
Information	[Type= discrete] [Format=character] [Missing=*]		
Definition	Last name or surname of the Person as it appears in DHIS-2 (or equivalent of national health information system)		
# petname: petname			
Information	[Type= discrete] [Format=character] [Missing=*]		
Definition	This is the Nickname, which is a familiar or humorous name given to a person to identify them within the household or locality		
# dob: dob			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=0 /-]		
Definition	Date of birth of the Person as recorded in the Health and Demographic Surveillance System (HDSS) dataset		
# sex: sex			
Information	[Type= discrete] [Format=numeric] [Range= 1-2] [Missing=*]		
Statistics [NW/ W]	[Valid=4115 /-] [Invalid=0 /-]		
Definition	The Gender of the person as recorded in the Health and Demographic Surveillance System (HDSS) dataset		
Value	Label	Cases	Percentage
1	Male	1447	<div><div></div></div> 35.2%
2	Female	2668	<div><div></div></div> 64.8%
9	Not Known or Not Mentioned	0	
Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.			
# nationalid: nationalid			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=4115 /-] [Invalid=0 /-]		
Definition	The National identifier like the National ID Card number or Maisha card number as recorded in the Health and Demographic Surveillance System (HDSS) dataset		
# hdssid: hdssid			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=4115 /-] [Invalid=0 /-]		
Definition	This is the unique identifier allocated to every individual or person in the dataset. This is different from the national identifiers and are local to the HDSS datasets.		
# hdsshid: hdsshid			
Information	[Type= discrete] [Format=character] [Missing=*]		



**File : synthetic\_hdss\_v3**# **hdsshhid:** hdsshhid**Statistics [NW/ W]** [Valid=4115 /-] [Invalid=0 /-]**Definition** This is the unique household identifier (number) allocated to every household in the dataset where a group of individuals or the persons reside.

File : synthetic_facility_v3			
# recnr: recnr			
Information	[Type= continuous] [Format=numeric] [Range= 2-2903] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	A sequential number to identify every record in the dataset uniquely		
# firstname: firstname			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	First name of the Person as it appears in the health facility (or equivalent of national health information system)		
# lastname: lastname			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	Last name or surname of the Person as it appears in the health facility (or equivalent of national health information system)		
# petname: petname			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	This is the Nickname, which is a familiar or humorous name given to a person to identify them within the household or locality		
# dob: dob			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=0 /-]		
Definition	Date of birth of the Patient as recorded in the Health FACility or the Health Center		
# sex: sex			
Information	[Type= discrete] [Format=numeric] [Range= 1-2] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	The Gender of the person as recorded in the Health Facility or the Health Center		
Value	Label	Cases	Percentage
1	Male	1024	35.3%
2	Female	1878	64.7%
9	Not Known or Not Mentioned	0	
Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.			
# nationalid: nationalid			
Information	[Type= discrete] [Format=character] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	The National identifier like the National ID Card number or Maisha card number as recorded in the Health Facility dataset		
# patientid: patientid			
Information	[Type= continuous] [Format=numeric] [Range= 2069-4970] [Missing=*]		
Statistics [NW/ W]	[Valid=2902 /-] [Invalid=0 /-]		
Definition	This is the unique identifier allocated to every patient during outpatient or in-patient visits in the health facility dataset. This is different from the national identifiers and are local to the Health Facility datasets.		

## File : synthetic\_facility\_v3

# visitdate: visitdate

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=0 /-]
Definition	The date the patient visited the health facility