**Cohort Profile: The Kiang West Longitudinal Population Study (KWLPS) – a platform for integrated research and health care provision in rural Gambia.**

Branwen J. Hennig1,2\*, Stefan A. Unger1,3\*, Bai Lamin Dondeh1, Jahid Hassan1, Sophie Hawkesworth1,2,4, Landing Jarjou1, Kerry S. Jones4, Sophie E. Moore1,2,4, Helen M. Nabwera1, Mohammed Ngum1, Ann Prentice1,4, Bakary Sonko1, Andrew M. Prentice1,2§, Anthony J. Fulford1,2\*

**Supplementary materials - contents**

1. Abbreviations in main text and Supplementary materials
2. Supplementary Figures 1-3
3. Kiang West Longitudinal Population Study (KWLPS) Databases, additional information
4. Kiang West Demographic Surveillance System (KWDSSS), additional information
5. Keneba Electronic Medical Records System (KEMReS), additional information
6. Keneba Biobank, additional information
7. Demography Search using Bayes (DSUB) algorithm for ID searches
8. References
9. **Abbreviations in main text and Supplementary materials**

DoB date of birth

DSUB demography search using Bayes

EC ethics committee

EMRs electronic medical records

ID identification

KDSG Keneba Database Steering Group

KEMReS Keneba Electronic Medical Records System

KWDSS Kiang West Demographic Surveillance system

KWLPS Kiang West Longitudinal Population Study

LSHTM London School of Hygiene & Tropical Medicine

MRC Medical Research Council

NCD non-communicable disease

SCC Scientific Coordinating Committee

y year(s)

WKNO West Kiang Number (unique ID)

1. **Supplementary Figures**

**Supplementary Figure 1.** Population pyramid of the Kiang West Longitudinal Population Study cohort (Data from 2013).



**Supplementary Figure 2.** Overall number of clinic presentations at MRC Keneba per month January 2010 – July 2014 for all age groups and residents in Kiang West.



**Supplementary Figure 3.** Stages of data collection within KEMReS. A similar structure exists for antenatal/postnatal visits and women seen by the midwives with review by a doctor if required.



1. **Kiang West Longitudinal Population Study (KWLPS) Databases, additional information**

Maintaining the integrity of the linkages within and between all KWLPS databases as well as project databases is critical in a database of this complexity, especially one used for scientific research where lost links and stray fragments can seriously distort results. In order to achieve this we impose basic constraints as outlined in the main manuscript. In addition we allow KWDSS data to be recorded from other sources. In particular, births and pregnancies may be derived from the maternity arm of KEMReS since the three-monthly KWDSS round is too infrequent for studies of pregnant women or neonates. This includes data from MRC Keneba antenatal health clinics and data gathered by the antenatal trekking team covering the whole of KW. Midwives record new pregnancies and births and the KWDSS database is updated and infants’ ID numbers assigned automatically. Similarly, information is rectified when patients present to the general Keneba clinic. This ensures that dates of birth are recorded accurately and up-to-date data on individuals is available from a single database table without needing to wait for the next KWDSS round.

Further developments planned to extend the existing infrastructure elements are a tracking system for all biological samples (already in place for Keneba Biobank) and creating a common data table for each laboratory analyser used at MRC Keneba.

All databases at MRC Keneba are backed daily and stored in a fireproof safe. Back-up copies of the database are also sent offsite to MRC Fajara (The Gambia) and MRC Head Office (UK) at regular intervals.

The genetic database contains information on candidate gene variation, Infinium HumanExome and Infinium HumanMethylation450 array data as well as candidate gene methylation data on between 130-3500 individuals and is currently held at London School Hygiene and Tropical Medicine (LSHTM).

1. **Kiang West Demographic Surveillance System (KWDSS), additional information**

All residents in the Kiang West district captured by the KWDSS are assigned a unique ID number, the WKNO. The WKNO includes a check-letter critical to minimise ID errors: any common typographic error (e.g. single-digit substitutions, reversal of adjacent digits etc.) will result in a number whose check-letter differs from that of the original and hence will be invalid. Throughout the database attempts to enter records with invalid ID numbers are always blocked.

It is often necessary to find the unique ID number of an individual based on limited information. This is not always a straightforward task in a community in which many people share the same name, name spellings are very variable, some people do not know their date of birth and there is no formal address system. For this purpose we devised a search utility, which we termed ‘Demography Search using Bayes’ (DSUB) algorithm to search the KWDSS database for individual(s). The user may input whatever is known of the individual’s given and surnames, date of birth or age estimates, parents’ names, village and/or compound of residence; the programme outputs the best matches. For further details see below.

Since there is no formal postal address system in the region, villages and compounds are also assigned a number, their GPS coordinates recorded and the villages mapped.

DOB is recorded and confirmed through several routes including: registration at birth, village recorders, infant welfare cards, questionnaires checking against historical events (e.g. measles epidemic, independence etc.).

KWDSS uses two main types of data tables - “constant” and “episode”. Constant tables are mostly used to record unchanging data. The most important such table is the Individuals’ table, which records individuals’ WKNO, name, sex, date of birth, self-reported ethnicity and parents’ ID numbers. There are others for compounds (recording names, current heads of compound, GPS coordinates) and villages (recording names, GPS coordinates and maps) and others recording alternative spellings of first and second names. The most important episode table records the residency episodes, recording intervals of time when an individual was living in a particular compound, i.e. start and stop dates, location and the events (e.g. birth, death, migration) that began and ended the episode. There are simpler episode tables for pregnancies and marriages. No individual is allowed more than one residency, one pregnancy and one relationship episode at a time, and the latter two only permitted for women over the age of 14 years.

1. **Keneba Electronic Medical Records System (KEMReS), additional information**
2. Other health care provision for Kiang West

Two other smaller government run health stations in the region exist (Figure 1). Karantaba (23.8 km to the West of MRC Keneba) primary health care clinic presently has three nurses and a midwife and has been supported over the last 35 years by MRC Keneba clinical staff. Kwinella primary health care station (36.6 km to the East of MRC Keneba) is presently staffed by four nurses and although situated outside of Kiang West it serves a small proportion of the KWLPS cohort. A permanent midwife is also stationed in Jiffarong and the villages Manduar and Kulikunda have government funded community health nurses (CHN). The closest hospital is in Bwiam (government run) that can be reached by boat or by road (1 ½ hours by car) (Figure 1). In The Gambia qualifications of nurses are graded according to the level and length of training1. State registered nurses and state registered midwives attend a 4-year course whereas state-enrolled nurses and state-enrolled midwives attend a 3-year course. Community health nurses and midwives focus on public health issues and participate in a 3-year training scheme. Most primary health care stations in The Gambia are not staffed with doctors1.

1. KEMReS set-up and access

KEMReS is written in Microsoft Visual Basic.NET (VB.Net) with a Microsoft SQL Server database backend. A stand-alone application was given priority as there were no other such system available in The Gambia and the system had to suit the existing setting and structure. A local area network (LAN) system was deployed similar to EMRs previously described2. VB.Net within LAN system allows extension for multi-user settings and data access and live data entry and concurrent data access from multiple access sites. Information can be communicated between multiple locations such as from the laboratory to the physician. VB.Net also allows flexibility in the areas of object-oriented programming, has easy syntax and semantics and provides a good platform for integration. New data fields required for future research projects or to improve patient management can be easily integrated.

KEMReS at the MRC Keneba health care clinic started recording patient attendances in December 2009 after an extensive development and training period of all clinical staff members. The clinic is currently staffed by four doctors, two midwives and ten nurses who are involved in clinical and research activities. One of the barriers to the introduction of use of Electronic Medical Records (EMRs) in LMIC settings is the lack of IT skills of clinical staff members3. Around 60% of our clinical staff members at MRC Keneba had never used a computer, with the remainder having IT skills not beyond simple writing skills.

Privacy and confidentiality is a major concern for EMRs4. Access to KEMReS is restricted to clinical staff members and only to entry screens relevant for position held and to those involved in the direct care to avoid uninformed and hence inferior patient care.

On presentation to the MRC Keneba clinic the patient is identified either with their KEMReS card containing the unique ID or using the DSUB search algorithm described below. The records office functions as the ‘Check In (Registration) and ‘Check Out’ (Check for a full data set for each patient encounter). In a medical emergency, paper copies of the main KEMReS data entry fields are used and entered by a trained administrator retrospectively. The number of administrators (or ‘super users’5), who are able to deal with making changes within the database is kept to a minimum to ensure confidentiality and data storage quality4.

1. **Keneba Biobank, additional information**

We introduced a custom-designed fully digitised database and sample tracking system in Keneba, which links in with the KWDSS, KEMReS, and all other phenotype databases held at MRC Keneba. The Biobank database system is used for all Biobank processes including:

1. generation of consent and call lists;
2. collection and logging of participant information on basic demographics, phenotypes and lab tests;
3. collection and logging of biological samples;
4. tracking of sample processing;
5. long- term storage at -80°C.

The sample handling and storage system is based on a 1D sample and 2D aliquot barcode tracking system, with multiple banked microtubes per participant (20 and 8 sample aliquots per >5 and <5y old, respectively) for ease of sample selection without freeze-thawing and facilitation of shipping, e.g. to analytical laboratories or collaborators, thereby ensuring the long-term integrity of the biological samples.

Clinical referral criteria were defined to identify urgent referrals based on malaria rapid test positivity (in <5y olds only) or severe high blood pressure (SBP>140 & DBP>90) observed during the field visit. Affected participants, as well as those who report feeling sick, are brought back to the MRC Keneba clinic on the same day for evaluation by a clinician. Non-urgent referrals are identified via a search function within the Keneba Biobank database, based on age-specific cut-offs for high blood pressure6, diabetes (fasting glucose >7.0mmol/l), and age-specific cut-offs for anemia7. Non-urgent referral cases are called to the MRC Keneba clinic within a few days for retesting and clinical evaluation during regular clinics (e.g. the weekly NCD clinic).

The Keneba Biobank study manual is available on request from the PI or the Head of Data at MRC Keneba.

The Keneba Biobank is part of the LMIC Biobank and Cohort Network (BCNet, <http://bcnet.iarc.fr/>)8, initiated by International Agency for Research on Cancer (IARC), together with the US National Cancer Institute - Center for Global Health (NCI-CGH) and other international partners.

1. **Demography Search using Bayes (DSUB) algorithm for ID searches algorithm**

It is often necessary to find the unique ID number of an individual based on limited information. Commonly, for instance, we need (i) to identify a new individual presenting at the clinic reception; (ii) when an individual moves between two compounds in the district, to link internal the separate reports of them leaving their original compound and entering the new one; (iii) to identify the parents of newly registered individuals; (iv) to find the KWDSS number of study subjects recruited prior to the advent of the KWDSS or (v) resolve data discrepancies, especially when they may involve an ID error. For this purpose we devised a search utility, which we termed ‘Demography Search using Bayes’ (DSUB) algorithm, that searches the KWDSS database for the individual(s) that best match whatever information the user provides. Briefly, the user may input whatever is known of the individual’s given and surnames, date of birth or age estimates, parents’ names, village and/or compound of residence. The programme outputs the best matches, listed in order of the posterior Bayes probability that we believe the match to be true. It allows for alternative name spellings (maintained in a table of alternative spellings), imprecision in the date of birth and the possibility of errors in the database.

More specifically, the ID search algorithm comprises the following components:

a) We have a database of N individuals referred to by the subscript i ∈ {1, 2, ... N}, for each of whom we have (possibly incomplete) personal data, and some fragmentary data on an individual (the “target”) whose ID number we wish to find.

X = vector of information on individual we are seeking;

Yi = vector of data (not necessarily same variables as in X) of personal details of the ith individual in the database.

Si = event that ith individual is the one we are seeking;

For each individual in the database we wish to find pr[Si | X], i.e. the probability that the ith member of the database is the individual we are seeking given the information, X, we know about him or her.

Ignore for now the possibility that the individual we are seeking is not in the database. From Bayes theorem:

 pr[Si | X] = pr[X | Si ]. pr[Si ] / pr[X]

 = pr[X | Si ]. pr[Si ] / (1)

b) We ignore information, such as age and sex, that might tell us how likely an individual in the database is to, say turn up at the clinic (we condition on much of this information anyway). In so doing we assuming that pr[Si] is independent of i, which allows us to cancel it from the equation:

 pr[Si | X] = pr[X | Si ] / (2)

So now all we need to find is pr[ X | Si ] ∀ I, i.e. the probability that should ith person in the database be the one we are seeking, they would present with the data, X, that we see.

For now, we consider xj and yij, the jth components of X and Yi, separately. We denote the corresponding random variable with a capital: Xj and Yij.

c) Consider first the simple case in which xj and yij are known accurately (if known at all) and correspond exactly.

 ⎧ 1 if (xj = yij and yij is known) or (xj is unknown)

 pr[Xj=xj | Si] = ⎨ 0 if xj ≠ yij and yij is known (3).

 ⎩ pr[Xj=xj] if yij is unknown.

[Note - we essentially omit the jth variable altogether if xj is missing. Setting the pr[Xj=xj|Si]=1 achieves this.]

d) More realistically we allow for some error, i.e. there is a possibility that yrecij, the value recorded in the database, is incorrect. We introduce the (predetermined) probability pr[Yrecij ≠ Yij]= pr\_errj. Then, in general:

 pr[Xj=xj | Si] = (1-pr\_errj).pr[Xj=xj | Si ∩ Yrecij = Yij] + pr\_errj.pr[Xj=xj | Si ∩ Yrecij ≠ Yij] (4)

Specific cases, in which xj and yij are categorical, continuous and when there are several values to match (e.g. names) are considered in 5, 6 & 7 below.

e) Consider the case when xj and yij are unordered and categorical. Si => Xj=Yij

Therefore (for simple, not “multiple”, matches):

 yrecij = xj => pr[Xj=xj | Si ∩ Yrecij = Yij]=1 and pr[Xj=xj | Si ∩ Yrecij ≠ Yij]=0.

Likewise:

 yrecij ≠ xj => pr[Xj=xj | Si ∩ Yrecij=Yij]=0 and pr[Xj=xj|Si ∩ Yrecij≠Yij]=pr[Xj≠yrecij].

Equation (4) then reduces to:

 ⎧ 1 if xj unknown

 ⎪ 1 - pr\_errj if yrecij = xj

 pr[Xj=xj | Si] = ⎨ (5)

 ⎪ pr\_errj. pr[Xj=xj]/(1 - pr[yrecij]) if yrecij ≠ xj

 ⎩ pr[Xj=xj] if yrecij unknown

We can readily estimate pr[Xj=xj] and pr[yrecij] from the proportions of individuals in the population who have these values for the jth variable.

f) Consider now the case when xj and yij are continuous and known with limited precision. We require estimates of the standard deviation associated with each level of precision. For KWDSS data, this only applies to the date of birth (DoB) for which we define three levels of precision tabulated below.

|  |  |  |
| --- | --- | --- |
| **precision level** | **description** | **Standard deviation** |
| D | nearest day | 2 days |
| M | nearest month | 30 days |
| Y | nearest year | 15% of the age when the DoB was first recorded |

The probability pr[xj | Si] requires both the distribution of difference between xj and yrecij (both of which should be measured in years), which we assume to be normal with mean=0 and variance = sd2(xj)+sd2(yrecij), and f[x] the distribution density of values in the database.

 ⎧ 1 if xj unknown

pr[Xj=xj | Si] = ⎨ (1-pr\_errj). 2\*Φ[-|xj-yrecij|/√{sd2(xj)+sd2(yrecij)}] + pr\_errj.f[xj] if yrecij known

 ⎩ f[xj] if yrecij unknown

 (6)

In order to calculate f[xj] we smooth the annual frequencies (ignore seasonality of birth frequency) and standardise to obtain a proper density function.. The smoothing is achieved simply by estimating the mean frequency for each decade and interpolating linearly. Note also, f[xj] needs to reflect the precision of the DOB: it should take the value f[xj] for precision level Y, f[xj]/12 for M and f[xj]/365 for D.

The Logistic rather than Normal distribution is used to calculate these probabilities since it provides a very close approximation (ref) and its distribution function, Φ[.|, is much easier to calculate.

g) Multiple matches (names).

1. Suppose an individual has several names in the main database, i.e. a set of names we will denote {*yrecij*}. Define a match as *xj* ∈{*yrecij*}. There are now two reasons why (5) does not apply. First, the *pr*[*Xj=xj* | *Si*] ≠1-*pr\_errj* even when there is a match; one of the other names might be given. Second, the probability when there is no match becomes tedious to compute. We must also redefine the error slightly to mean that there is an error in the name given as target.

To deal with the first we define two types of first names: formal ones and “aka”s. Some people may have several formal names, but would normally use one of them – though we don’t necessarily know which – unless they used their aka. We have no idea what proportion of the time the aka or formal name would be given so we (initially) estimate that to be 50:50. Thus, if there is an aka, we divide the probability by 2: *pr*[*Xj=xj* | *Si*] =(1-*pr\_errj*)/2, otherwise *pr*[*Xj=xj* | *Si*] =1*-pr\_errj*.

The second problem is that when there is a mismatch, i.e. when xj∉{yrecij}, *pr*[*Xj=xj* | *Si*] = *pr\_errj.pr*[*Xj=xj* | *Xj*∉{*yrecij*}], and *pr*[*Xj*=*xj* | *Xj*∉{*yrecij*}] is tedious to estimate. However it is only slightly different from the unconditional probability of obtaining xj, as required when {*yj*} is unknown.

 (7)

Even the precise calculation of *pr*[*Xj=xj*] is impossible, requiring, as it does, the likelihood that each name an individual has will be given as a target. Instead we estimate this as the relative frequency of the name among all the names given. Thus, where nnamei is the number of names the ith individual has and *I*(*xj* ∈{*yrecij*}) is an indicator function taking the value 1 or 0 depending on the truth of the argument (i.e. whether *xj* ∈{*yrecij*}), we estimate:

 (8).

This can be readily calculated before hand for each name.

Furthermore, names are usually sex-specific so we calculate a separate table of name probabilities for boys and girls. We maintain a list of first names used in Kiang West and associated sex for each. Some names are used for either sex; these are coded “E” (for either) and the probability calculated as the average of the boys and girls probabilities.

1. Multiple target names. We currently only allow one first name in the target information.

h) When combining the probabilities for each component we ought, strictly, to allow for the dependence between the components of X:

 pr[X | Si] = pr[x1 | Si] . pr[x2 | Si ∩ x1] . pr[x3 | Si ∩ x1 ∩ x2] ... (9)

In practice, these conditional probabilities cannot be calculated easily. Instead we assume they are independent, i.e. , but impose some rules in order to ensure that this is at least approximately true. Some rules we apply are:

1. Only allow matching on the name OR id number of parents, not both.
2. Don’t match on sex if the individual’s first name is known (since this will essentially determine the sex, and when it doesn’t there is often doubt about whether the sex if correctly recorded anyway).
3. Only allow searches on compound if the village is specified: village+compound then becomes a single variable.
4. Only allow compound to be matched on one of compound name, compound head or compound number.
5. Don’t match on both individual’s and father’s second name (they are usually the same and, if not, there may be ambiguity).

Thus the list of variables to match on will be:

* individual’s first name OR sex (one variable);
* individual’s second name;
* mother’s first and second names (two variables) OR her number (one variable);
* father’s first name OR his number (one variable);
* residence: village and one of compound name, head or number (one combined variable) or just village if compound is unknown;
* dob (incorporating its precision);

Many extensions and refinements to this system are clearly possible but, crude as it is, the current system does perform remarkably well. It has two major advantages over a simple search: (a) it allows for data errors and (b) it reports the probability of the matches found. It is important to take account of posterior probabilities before accepting a match. It is very helpful in revealing that there are many equally good matches, the temptation otherwise might be to accept the first found. It is also possible that one match may appear much better than any other but still be uncertain due to the number of near matches the cumulative probability of which may be considerable.

1. **References**

1. Risk R, Naismith H, Burnett A, Moore SE, Cham M, Unger S. Rational prescribing in paediatrics in a resource-limited setting. *Arch Dis Child*. 2013;503–509.

2. Douglas GP, Deula R a, Connor SE. The Lilongwe Central Hospital Patient Management Information System: a success in computer-based order entry where one might least expect it. *AMIA Annu Symp Proc*. 2003;833.

3. Curioso WH. New Technologies and Public Health in Developing Countries: The Cell PREVEN Project. *LEA’s Commun Ser*. 2006;393–403.

4. Mandl KD, Szolovits P, Kohane IS. Public standards and patients’ control: how to keep electronic medical records accessible but private. *BMJ*. 2001;**322**(7281):283–287.

5. Baron RJ, Fabens EL, Schiffman M, Wolf E. Electronic health records: Just around the corner? Or over the cliff? *Ann Intern Med*. 2005;**143**(3):222–226.

6. Oussama M, Mohammed S. Clinical Guidelines for the Management of Hypertension. WHO; 2005. p. 14–30, 48–51.

7. WHO. Pocket book of hospital care for children: guidelines for the management of common childhood illnesses. 2013;1–483.

8. Mendy M, Caboux E, Sylla BS, Dillner J, Chinquee J, Wild C. Infrastructure and Facilities for Human Biobanking in Low- and Middle-Income Countries: A Situation Analysis. *Pathobiology*. 2015;**81**(5-6):252–260.