International Journal of Population Data Science

Journal Website: www.ijpds.org

Establishment of a birth-to-education cohort of 1 million Palestinian refugees using electronic medical records and electronic education records

Zeina Jamaluddine[1,2], Akihiro Seita[3], Ghada Ballout[3], Hussam Al-Fudoli[3], Gloria Paolucci[3], Shatha Albaik[3], Rami Ibrahim[3], Miho Sato[2], Hala Ghattas[4,5], and Oona M. R. Campbell[1,*]

Submission History

Submitted:	14/04/2023
Accepted:	04/09/2023
Published:	24/10/2023

[1]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom
[2]School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan
[3]United Nations Relief and Works Agency for Palestinian Refugees in the Near East, UNRWA headquarters, Amman, Jordan
[4]Department of Health Promotion, Education, and Behavior, University of South Carolina, Columbia, South Carolina, USA
[5]Center for Research on Population and Health, Faculty of Health Sciences, American University of Beirut, Beirut Lebanon

Abstract

Introduction

By linking datasets, electronic records can be used to build large birth-cohorts, enabling researchers to cost-effectively answer questions relevant to populations over the life-course. Currently, around 5.8 million Palestinian refugees live in five settings: Jordan, Lebanon, Syria, West Bank, and Gaza Strip. The United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA) provides them with free primary health and elementary-school services. It maintains electronic records to do so.

We aimed to establish a birth cohort of Palestinian refugees born between 1st January 2010 and 31st December 2020 living in five settings by linking mother obstetric records with child health and education records and to describe some of the cohort characteristics. In future, we plan to assess effects of size-at-birth on growth, health and educational attainment, among other questions.

Methods

We extracted all available data from 140 health centres and 702 schools across five settings, i.e. all UNRWA service users. Creating the cohort involved examining IDs and other data, preparing data, de-duplicating records, and identifying live-births, linking the mothers' and children's data using different deterministic linking algorithms, and understanding reasons for non-linkage.

Results

We established a birth cohort of Palestinian refugees using electronic records of 972,743 live births. We found high levels of linkage to health records overall (83%), which improved over time (from 73% to 86%), and variations in linkage rates by setting: these averaged 93% in Gaza, 89% in Lebanon, 75% in Jordan, 73% in West Bank and 68% in Syria. Of the 423,580 children age-eligible to go to school, 47% went to UNRWA schools and comprised of 197,479 children with both health and education records, and 2,447 children with only education records. In addition to year and setting, other factors associated with non-linkage included mortality and having a non-refugee mother. Misclassification errors were minimal.

Conclusion

This linked open birth-cohort is unique for refugees and the Arab region and forms the basis for many future studies, including to elucidate pathways for improved health and education in this vulnerable, understudied population. Our characterization of the cohort leads us to recommend using different sub-sets of the cohort depending on the research question and analytic purposes.

Keywords

electronic records; data linkage; mother child; Palestinian refugees; health records; education records; refugee birth cohort

*Corresponding Author:
Email Address: oona.campbell@lshtm.ac.uk (Oona M. R. Campbell)

https://doi.org/10.23889/ijpds.v8i1.2156

Introduction

Refugees and urban-poor populations remain under-studied globally because their unstable living circumstances make them difficult to research, especially longitudinally. The Arab region has few longitudinal cohorts [1], little research on refugees or the urban-poor, and limited research using individual-level electronic records at a large scale.

Large longitudinal studies are enormously beneficial in elucidating factors shaping human capital, including health and educational outcomes [2, 3]. The use of existing electronic records to build large birth-cohorts offers a cost-effective alternative to traditional birth cohorts, enabling researchers to answer questions relevant to populations over the life-course. Linked data make more information available, allowing analyses in different domains, for instance, understanding the effects of ill-health on educational attainment. Recently for example, linked administrative data have been used to model disease patterns and to examine factors associated with COVID-19 infection and related deaths to inform timely policy changes [4, 5].

Electronic data present challenges in terms of data-capture and linkage [6]; it is important to detect the extent of errors including misclassification, temporal data changes, missing data, and duplicated records and to identify the population included and excluded. Linking electronic data adds further challenges depending on the methods used for linkage (deterministic or probabilistic methods), the presence of duplicated records (causing additional linkage error), estimation of error rates (with challenges in obtaining a gold standard), and identification of the population (understanding who does and does not link) [6–8].

Population and settings

Palestinian refugees include all descendants of Palestine refugee males, who are "persons whose normal place of residence was Palestine during the period 1 June 1946 to 15 May 1948, and who lost both home and means of livelihood as a result of the 1948 conflict". Palestinian refugees comprise 20% of the global refugee population and have experienced displacement and marginalisation since 1948 [9]. Currently, around 5.8 million Palestinian refugees live in 58 camps and multiple informal gatherings in five settings: Jordan, Lebanon, Syria, West Bank, and Gaza Strip (representing an estimated 45% of all Palestinians) [9, 10]. The United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA) is responsible for providing free primary health and elementary-school services to the refugees [10], and runs 140 health centres and 702 schools to do so [11]. UNRWA also supports Palestinian refugees' access to secondary and tertiary health-care services via a partial reimbursement scheme.

Access to UNRWA services differs by setting (Box 1). In 2021, UNRWA recorded 3,090,084 refugees accessed their health services, indicating not all 5.8 million Palestinian refugees are UNRWA service recipients [12]. Some, particularly the better-off, may use alternative services in the host communities [12, 13], yet others may use a mix of UNRWA and other service providers. In 2021, UNRWA recorded that 526,646 students attended their schools; as with healthcare [11], not all Palestinian children enrol in UNRWA schools, and some use host-country public or private schools.

UNRWA has consistently invested in record- keeping, and now maintains electronic administrative databases to provide its health and education services, namely an electronic health records system (E-health) and an Education Management Information System (EMIS).

E-health was developed in 2010 as a web-based, patient-centred digital system to manage UNRWA's increasing workload and to improve the quality of its health care provision [14]. E-health started gradually in clinics and was updated in 2013 and 2017. EMIS was launched in the 2016/2017 school year to manage education data in UNRWA schools and improve overall educational quality. Both systems include identification numbers (IDs) which allow for deterministic linkage.

Aim

We aimed to build a live birth-cohort to enable us to explore the effects of risk factors and exposures in pregnancy (e.g., previous obstetric history, complications in pregnancy, and pollution, temperature, and conflict), and of factors recorded via the obstetric record, (e.g., pregnancy outcome, gestation, birthweight, multiples, and mode of delivery) on adverse health and educational outcomes among children.

We identified a group of women eligible to access UNRWA services with a pregnancy that ended from 2010-2020. For the subset with live births, we aimed to link information from mothers' obstetric records to UNRWA child health records and education records to create a live-birth cohort, and to describe some of its characteristics.

Methods

To create the cohort, we 1) examined IDs and other data, 2) prepared the data, de-duplicated records, and identified live-births 3) linked the mothers' and children's data using different deterministic linking algorithms, and 4) clarified reasons for non-linkage.

Examining IDs and other data

A cohort of refugees from E-health and EMIS

All records of pregnancies that ended between 1 January 2010 and 31 December 2020 (whether they resulted in live birth, early foetal death, stillbirth, miscarriage) were extracted from E-health, as were health and education records of children born in the same period in Jordan, Lebanon, Syria, West Bank and Gaza.

All the health information was stored in the E-health system while all the education data was stored in the EMIS system.

Figure 1 shows the key variables extracted from each of the three dataset: (1) mother E-health dataset including mother information, mother antenatal care (ANC) visit, mother obstetric records, (2) child E-health dataset including child information, child health data (including immunisation, growth monitoring, motor development, physical examination, outpatient visits, and laboratory results), (3) child EMIS

Box 1: Political, social, health and education system context of Palestinian refugees

	Country (setting) where Palestinian refugees are located				
	Jordan	**Lebanon**	**Syria**	**West Bank**	**Gaza**
Number of registered refugees reported by UNRWA in 2021 [11, 12]	2,334,789	482,676	575,234	883,950	1,516,258
Estimated percentage of total national population that are refugees (World Bank population data in 2021 [12, 15])	21%	9%	3%	30%	78%
Political/Social context	Most Palestinians have Jordanian nationality since 2009. Use of Jordanian government services permitted.	Palestinians' right to work & access to government services is severely constrained. Not eligible to use Lebanese public primary healthcare or schools.	Massive internal displacement since 2011. 137,234 Palestinians in Syria fled to Lebanon & Jordan; an estimated 438,000 remain.	Dual systems for Israeli settlers & Palestinians, restricting Palestinian rights and travel. Eligible to use public Palestinian Authority services.	Blockade & travel restrictions. Eligible to use public Palestinian Authority services.
Health and education context	UNRWA co-finances hospitalisation services.	UNRWA provides secondary school education. UNRWA co-finances hospitalisation.	Starting in 2011, UNRWA services affected by conflict. UNRWA co-finances hospitalisation.	Multiple checkpoints restricting access. UNRWA co-finances hospitalisation.	UNRWA co-finances hospitalisation services.
Number of UNRWA health centres in 2021 [11]	25	27	23	43	22
Number of UNRWA schools in 2021 [11]	161	65	102	96	278
Estimated pregnant Palestinian refugees using UNRWA antenatal care services in 2022 [16]	35%	63%	42%	54%	73%
Pregnant women using UNRWA antenatal care at least once with 4 or more antenatal visits in 2022 [16]	81%	75%	55%	90%	98%
Deliveries by trained personnel in 2022 [16]	100%	100%	100%	100%	100%
Children aged 12 months old receiving all vaccine immunisation (BCG, IPV, Poliomyelitis, DPT, Hepatitis B, Measles, Hib) in 2022 [16]	99%	97%	98%	100%	99%

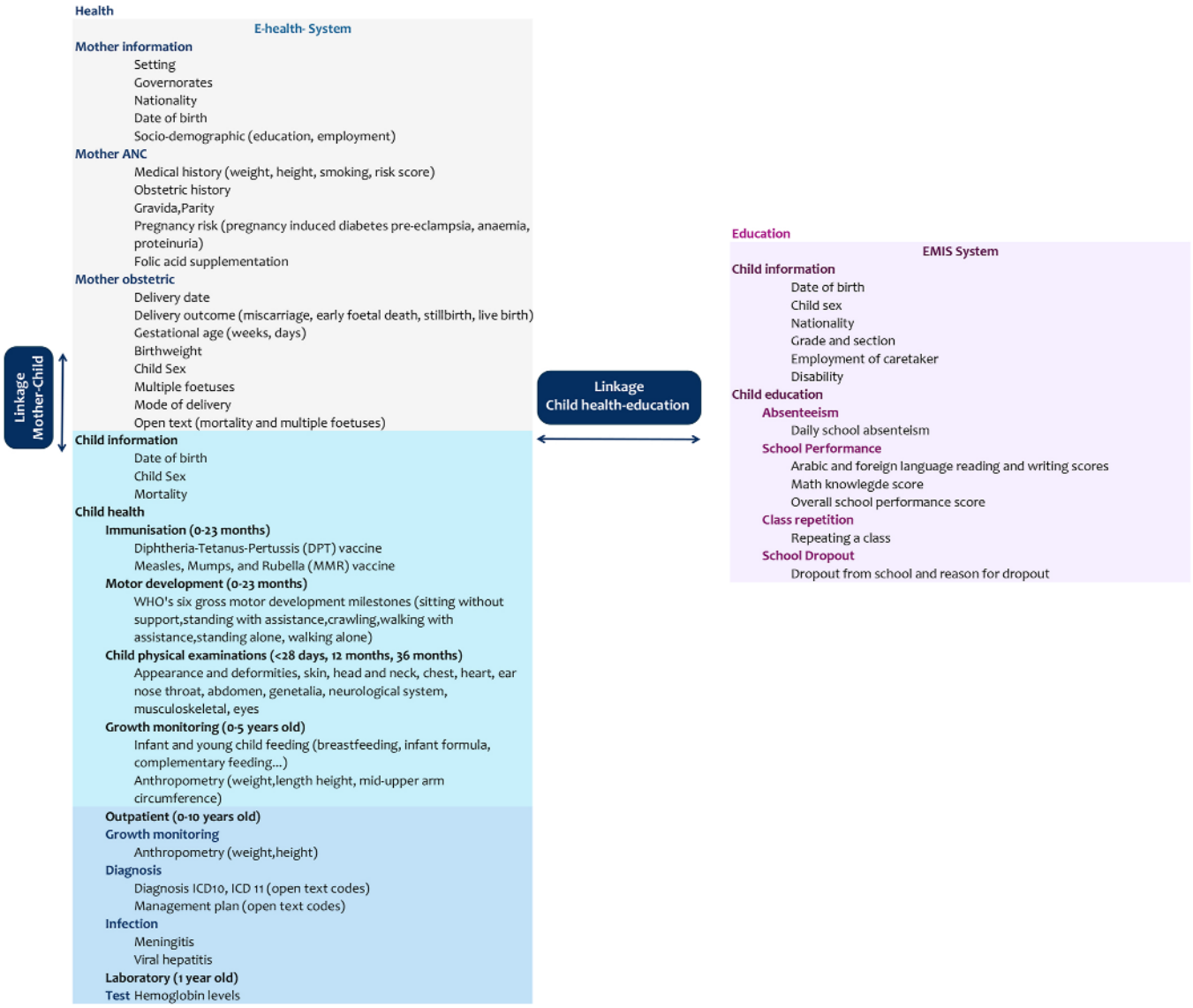
dataset including child education information, and child education.

Data from the mother's information included sociodemographic information and IDs with which to link mothers to their children. Mother's ANC records included women's medical history, reproductive health, and ANC received. Data from the mother's obstetric records included date of delivery, delivery outcome (live birth, stillbirth, early foetal death, miscarriage), multiple foetuses (twins/triplets/quadruplets), birthweight, gestational age, place of delivery, sex of live births, and mode of delivery. UNRWA partially covers childbirth costs, so neonatal information is gathered from the hospital records at billing and entered into the system after delivery as part of the women's obstetric records. Active surveillance of pregnancy outcomes for women who sought ANC takes place, with a call-back mechanism in case no pregnancy outcome is recorded.

UNRWA's primary care model provides for children to routinely undergo specific preventive measures, and it collects child health data on these accordingly, including immunisation as per host country schedules, growth monitoring (ages 0 to 59 months), motor development (ages 0 to 23 months), and periodic physical examinations (for new-borns and at 12 and 36 months). In 2017, they introduced mandatory screening for anaemia at 12 months. When care is sought for children who are ill, there may be additional outpatient records or laboratory-test results. This is an open cohort, for the children E-health records we extracted data from 01 January 2010 until 14 September 2021 (the day of the extraction) for the linkage.

Elementary school enrolment is mandatory (and free) from 6 years of age in all settings. Children born in 2010 would have reached age 6 and entered Grade 1 beginning in the 2016/17 academic year, with subsequent birth years entering school in the following years. Extraction of the education data was done in a yearly basis with total of 5 academic years extracted. Data

Figure 1: Variables extracted from maternal, child health records and child education records (grey mother E-health dataset, blue child E-health dataset, purple child EMIS dataset)



from EMIS captures information on students as they enter Grade 1 and progress in school from one year to the next. Extracted data included student characteristics, absenteeism, school performance, special education, class repetition and school drop-out.

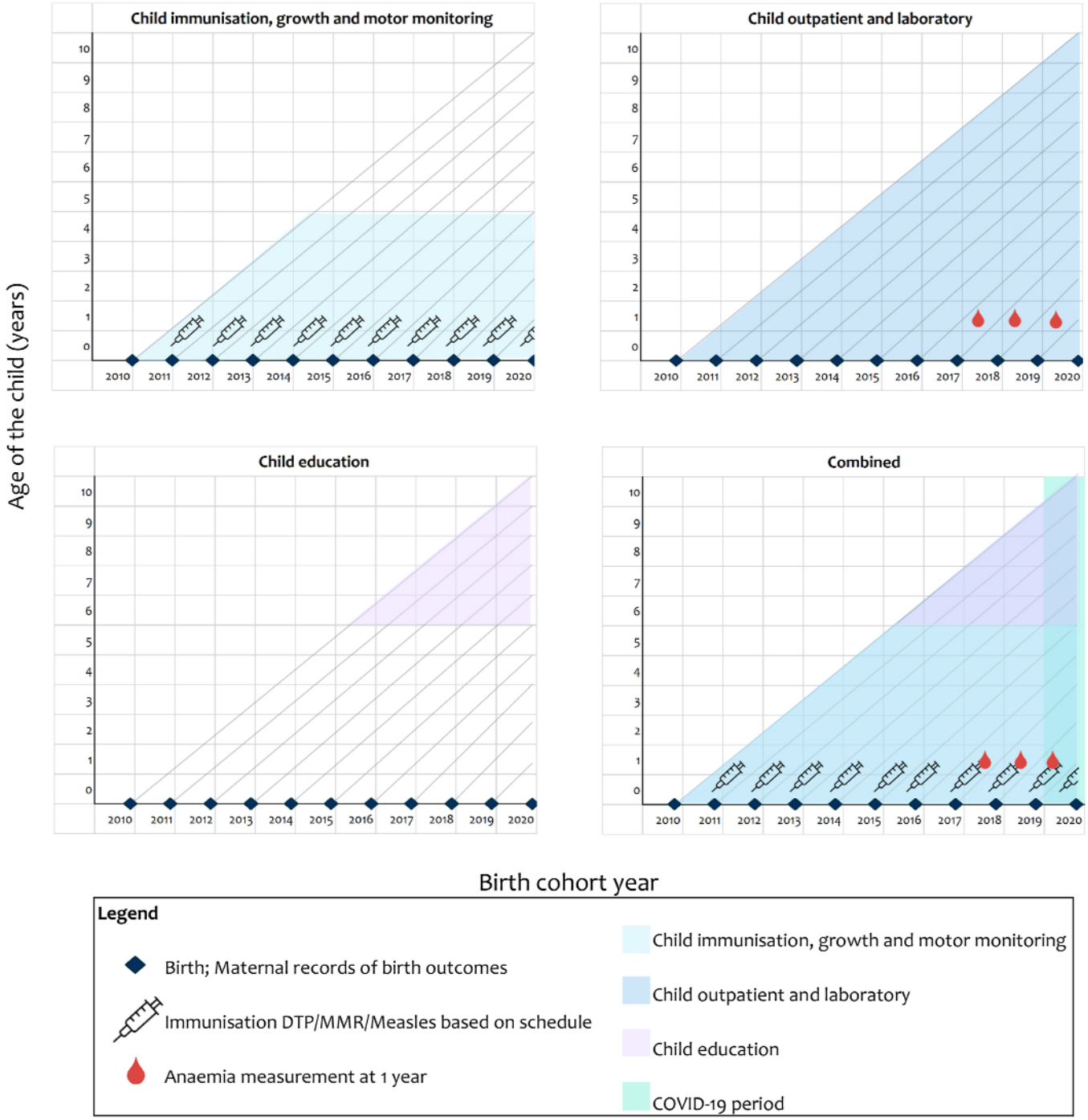
The Lexis diagram (Figure 2) indicates the different component datasets, and the time points when specific records could be accessed from E-health and EMIS to contribute to the birth cohort. The cohort (year of birth) is depicted as a blue diamond along the x-axis and the age of the child is on the y-axis. The shaded areas indicate the availability of data based on the cohort, age of the child, and the different types of records. For example, all children are expected to have records of routine preventive child health care (immunisation, growth, and motor monitoring records) up to 5 years of age, and education records starting at 6 years old. Haemoglobin level measurement for children at 1 year old (indicated using a red droplet) started in 2017. Outpatient and laboratory records are available at all ages (and into adulthood) for those needing these services.

Identifying IDs in different records

Six different IDs could potentially be used for linkage. The E-health system generates a unique Mother Medical File Number (MMFN) for each mother and a unique Child Medical File Number (CMFN) for each child. UNRWA also generates a unique refugee registration ID (RRIS) for each refugee and a family registration ID: MRRIS for mothers; CRRIS for their children; and FRRIS for families. As with birth registration, the CRRIS is generated when parents register their child in the system, so the first ID a child usually gets is the CMFN which is generated automatically by the E-health system when the child uses UNRWA services. In some cases, if the child was never taken to UNRWA services, the obstetric record might not have a CMFN. The CRRIS and FRRIS are also recorded in EMIS.

The different IDs available in the various mother, child health, and education records are shown in Figure 3, with pink and blue lines highlighting the IDs used to link across the various datasets. The obstetric records (column 3) include

Figure 2: Lexis diagram containing age of the child and source of the variables collected



data on neonatal outcomes without child-specific IDs. Child-specific IDs, the CMFN, are listed for each woman in the mother information records (column 1) showing all her children who have used UNRWA services; this is unlinked with neonatal outcomes.

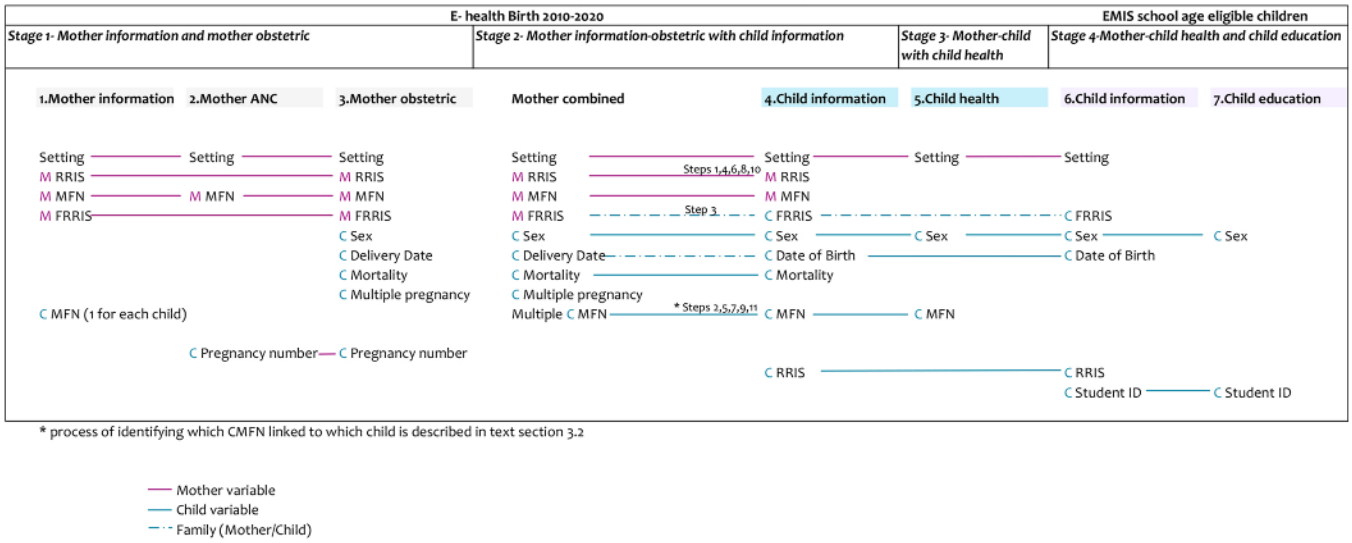
Data preparation, de-duplication of records and identification of multiple pregnancies

Data preparation involved cleaning specific open-text variables, selecting live births, distinguishing between multiples (twins, triplets, etc.,) and duplicated records, and removing the latter.

We cleaned open-text fields using text-mining tools to check different phrasings and spellings of "twins," "triplets," "quadruplets," and "multiples" (e.g., "tribblets") and of "death" (e.g., "died at 3 min") in English and Arabic. We then standardised the coding of these terms.

To link live births to children's health and education records, we excluded records of pregnancies that ended in miscarriage, early foetal death, or stillbirth, and records that were marked as training data or erroneous data (system-recorded errors). We then marked mothers' obstetric records with the same mother and the same delivery date as being either potential multiple pregnancies or duplicated records.

Figure 3: Steps in linkage of the different datasets (indicating the IDs used for linking in the different steps)



M RRIS	Mother refugee registration ID
M FRRIS	Mother family refugee registration ID
M MFN	Mother medical file number
C RRIS	Child refugee registration number
C FRRIS	Child family refugee registration ID
C MFN	Child medical file number
Steps	ID used for a specific step in the multi-step deterministic linkage

Data linkage using multi-step deterministic methods

The data linkage was conducted in four stages, linking: Stage 1) the mothers' information to the mother's ANC and the mother's obstetric outcomes; Stage 2) the mother's data from Stage 1 to the child information dataset; Stage 3) the dataset from Stage 2 to the child health datasets; and Stage 4) the information from Stage 2 to the child education dataset. The overall ID linkage stages are presented in Figure 3. When doing the linkage, we blocked on setting because each setting generates its own data (which are concatenated in UNRWA headquarters) and refugees rarely move across settings, and to help with managing a very large data set.

We ran the linkage process twice. The first time included only multiple pregnancies and duplicated records and aimed to distinguish duplicated records from multiple pregnancies based on the stages 1-4 described above. In some cases, data for multiples were entered as a single birth, with open text indicating the delivery resulted in twins, triplets, or quadruplets. We generated (and flagged) synthetic records for these missing multiples (additional twins, triplets, etc.,) based on the original obstetric record. This involved creating a record for each child we knew about to: (1) redress a limitation in the data structure as designed in E-health (2) allow us to have a more comprehensive dataset with proper denominator of live births, (3) retain information on maternal variables such as age and education, as well as ANC variables. These maternal and ANC attributes apply equally to all foetuses in a given pregnancy. It was only birthweight that is potentially incorrect. This information is clearly flagged in the dataset, and we record clear information to data users on how these variables can be used effectively.

Duplicated records were those with no mention of being a multiple, and where the records did not link to two or more different child IDs. We removed these duplicates leaving only one record, and then re-ran the linkage stages 1-4 with all the records (multiples and de-duplicated records, and all non-duplicate records). We removed the duplicated records that had missing data in one record entered in the birthweight measure as compared to the other records. Where duplicate records had contradicting information, we kept the last record of the duplicated records.

We calculated the percentage linking after adding synthetic records for missing multiples and removing duplicated records. We also assessed percentages after removing children with missing CMFNs (i.e., only keeping children that used UNRWA services).

Stage 1- linking mother's data sets

We first used the MMFN to link the mother's information with her ANC records and her obstetric outcomes. The mother's information includes a CMFN of all her children, without information on their birth order, date of birth, and sex. The mother's obstetric record lists the child's sex, delivery date, and multiple pregnancies, but does not have a child ID, for example, a CMFN.

Stage 2- linking mother information- obstetric with child information (from E-health)

In Stage 2, we merged the Stage 1 mother datasets (1–3) with the child information dataset (4) via 11 steps. In all steps, we blocked on setting to ensure this was identical in both mother records and child records. In steps 1 to 9, we linked

the singleton births, then in steps 10 and 11, we linked multiple births. Since the MRRIS is the most accurate ID (most used in UNRWA to refer to individual refugees), we used this first as the "best linkage," followed by linkage based on the CMFN.

Because the child's information had the date of birth and the CMFN, we could then match the date of delivery/birth upon linking to the CMFN from the mother's information, ensuring the obstetric record was given to the correct child. This approach worked for singleton or twins of discordant sex, but not for multiples of concordant sex.

In steps 1 to 3, we linked based on records having an identical month and year of delivery/birth, the same sex of the child, and the same MRRIS (step 1), same CMFN (step 2), and same FRRIS (step 3). In steps 4 to 5, we linked based on records having a delivery/birth date within plus or minus 90 days of each other and the same MRRIS (step 4), or the same CMFN (step 5). Then we allowed the delivery/birth date to be plus or minus 180 days and the same mother MRRIS (step 6), or the same CMFN (step 7). This was done to take into account the data entry errors in the delivery date or the date of birth. In steps 8 and 9, we removed the requirement for identical sex in steps 1 and 2, and linked based on MRRIS ID (step 8), and the same CMFN (step 9).

For twins, triplets, and quadruplets (multiples) we linked based on identical dates of delivery/birth and same mother MRRIS ID (step 10) and same CMFN (step 11).

Stage 3- linking mother-child information with child health

We used the CMFN to link the dataset from Stage 2 (which linked datasets 1-4) to the child's health records (dataset 5) including immunisation, growth monitoring, motor development, haemoglobin testing, outpatient visits, and laboratory results records.

Stage 4- linking mother-child health data with child education

Children from Stage2 who reached age 6 years or above were linked to EMIS datasets 6 and 7 based on the CRRIS, identical setting, sex, and month and year of birth.

Reasons for failure to link

We developed hypotheses about structural (legitimate) and other reasons for data to not link and tested these using a classification and regression (CART) decision tree approach [17] to identify groups at substantial risk of not linking. CART repeatedly separates data into two groups, one with high levels of non-linkage and one with low by testing different cut-off points (for example the different year of delivery/birth), and splits the data based on the best within-group homogeneity.

Information on mortality, mother's refugee status, year of delivery/birth, sex of the child, birthweight, and gestational age were available and were used to predict non-linkage. Mortality of the child (whether neonatal or infant or other) was included in the mother's obstetric records (thus this information is available in both linked and unlinked data). We also generated a low risk of mortality group (normal birthweight, term and singleton and not recorded as dead), recorded mortality, and a composite of low birthweight, preterm, or multiples without recorded mortality (as a measure of being a high risk of mortality that may not be recorded). Children of non-refugee mothers, but where the male parent is a refugee, are included in these datasets because they are eligible for UNRWA services.

We also ran a multivariable regression analysis looking at determinants of failure to link (Appendix).

The data cleaning, linkage and multivariable analyses were conducted using Stata software (StataCorp. Stata Statistical Software: Release 17. College Station, TX: StataCorp LL). The CART analysis was conducted using R software and rpart package (R Core Team, 2022, version 4.2.1 R Foundation for Statistical Computing, Vienna, Austria).

Results

Examining IDs and other data

From 1 January 2010 until 31 December 2020, a total of 1,158,354 pregnancy outcomes were extracted (Figure 4). For the linkage, we excluded 181 system-recorded errors as indicated in the open text and a total of 172,545 miscarriages, early foetal death, and stillbirth records and 181 system-recorded errors as indicated in the open text (Figure 4).

Data preparation and de-duplication of records

3,851 birth records were synthetically added when pregnancy outcomes were marked as multiples (twins, triplets, or quadruplets), but only one birth record was available. In some cases, we had one record indicating both death and multiple (for example "1 twin died while the other survived"), another record indicating this was synthetically added. A total of 45,095 records had at least one other record with the same Mother IDs and the same delivery date. We were able to distinguish 18,378 as multiple pregnancies (as noted in the open text variable), 9,981 as singleton records, and 16,736 as duplicated records. We dropped the latter.

This resulted in a total of 972,743 live birth records, born to women with an obstetric record, recorded in all five settings, of which 424,616 became eligible for school enrolment in the study time-period. From the child health records, in E-health, a total of 1,089,568 were extracted for the linkage. A total of 279,758 child education records were extracted from EMIS for the linkage. A total of 12,245 records mentioned death in an open text variable.

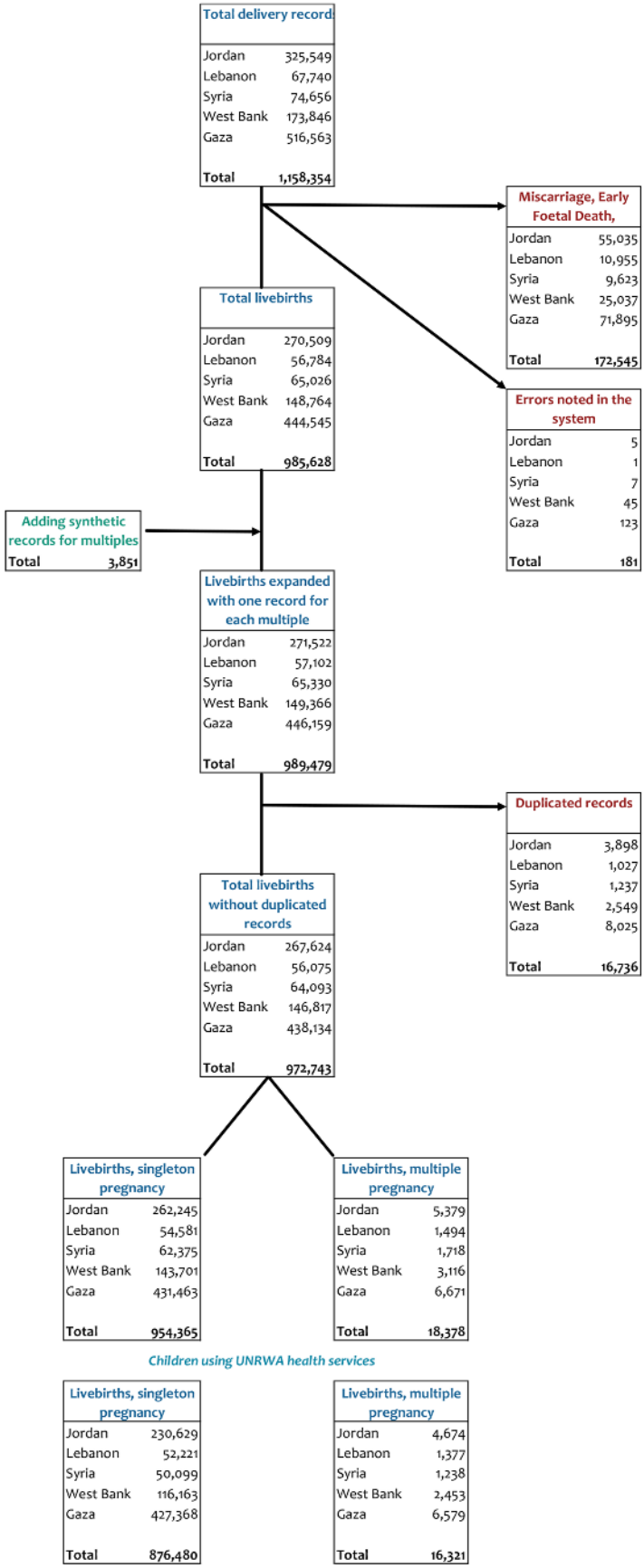
Data linkage

Deterministic linkage mother with child health records (Steps 1 to 11)

Linkage increased from 69% in step 1 to 83% in step 11 (Table 1). The percentage linkage between mother and child records improved from 73% in 2010 to 86% in 2020 (Figure 5). Gaza had the highest linkage, followed by Lebanon, Jordan, Syria, and the West Bank (Figure 5).

In 79,942 cases, there was no CMFN in the mother's records, most likely because their children did not use UNRWA services. By removing records with missing CMFN (in the

Figure 4: Data preparation and de-duplication of records



unlinked dataset) to make a birth cohort of UNRWA health service users (mother used UNRWA ANC or obstetric services and child had at least one record within the E-health) linkage improved to 91% overall, from 81% in 2010 to 94% in 2020. We refer to these as the "Stage 2 dataset" and the "Stage 2 dataset with children that used UNRWA health services".

Table 1: Mother-child linkage steps: matching requirements

Mother-child link Steps	Field/ Setting	Sex	Multiple/ Duplicated records	ID used	Date of birth/ Delivery date match	Numbers	Linkage (%) N = 972,743	Linkage (%) children using UNRWA health services N = 892,801
1	Exact	Exact	No	M RRIS	Month and Year	674,968	69%	76%
2	Exact	Exact	No	C MFN	Month and Year	708,586	73%	79%
3	Exact	Exact	No	F RRIS	Month and Year	732,802	75%	82%
4	Exact	Exact	No	M RRIS	±90 days	751,110	77%	84%
5	Exact	Exact	No	C MFN	±90days	752,916	77%	84%
6	Exact	Exact	No	M RRIS	±180 days	773,816	80%	87%
7	Exact	Exact	No	C MFN	±180 days	774,556	80%	87%
8	Exact		No	M RRIS	Month and Year	787,608	81%	88%
9	Exact		No	C MFN	Month and Year	788,239	81%	88%
10	Exact	Exact	Multiple	M RRIS	Month and Year	811,221	83%	91%
11	Exact	Exact	Multiple	C MFN	Month and Year	811,871	83%	91%

M RRIS Mother refugee registration ID.
C MFN Child medical file number.
F RRIS Family refugee registration ID.

Figure 5: Percentage of mother-child linkage over time (a) overall (b) Stage 2 by setting and (c) Stage 2 children that use UNRWA services

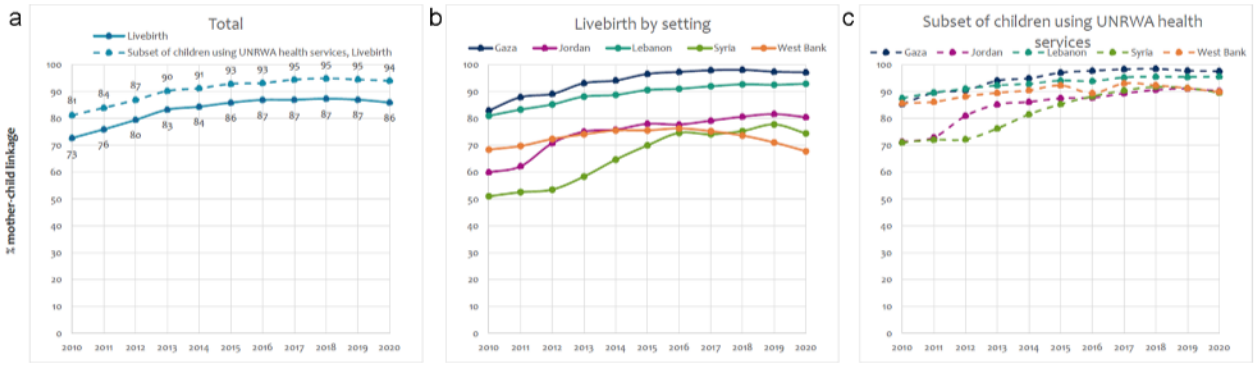


Figure 6 shows the percentage contributed by each linkage step per year. The percentage linking in steps 1 and 2 increased over time. Errors in the identical recording of the delivery/birth date were allowed for steps 4/5 (± 90 days) and step 6/7 (± 180 days); they decreased over time. Errors in recording sex (steps 8/9) were small and consistent over time. The percentage of multiples linking (steps 10/11) was the same across all years.

Deterministic linkage of mother-child health records with education records

The dataset from Stage 2 was linked to child health records (Stage 3) and education records (Stage 4). Linkage of the Stage 2 dataset of children that use UNRWA services to the child health records was extremely high at 98% (Figure 7).

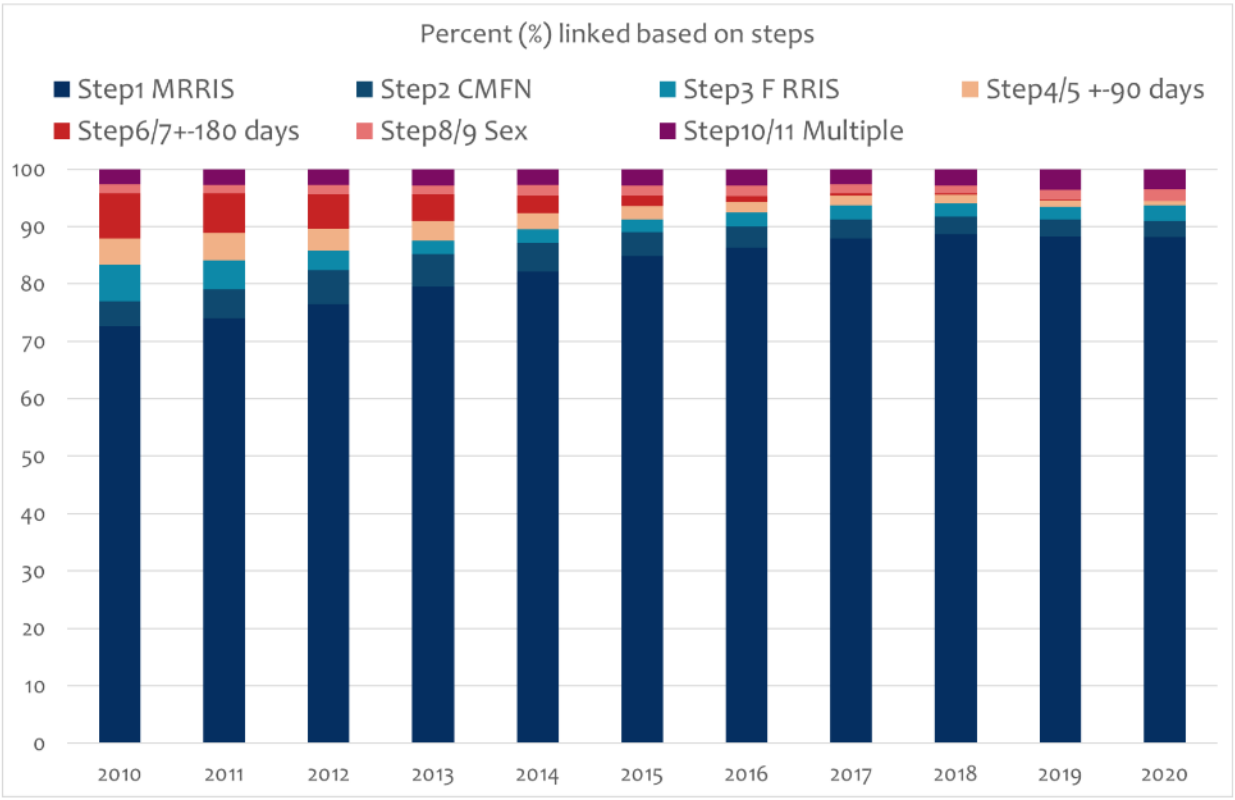
The live birth dataset had 424,616 records of children at an eligible age for school enrolment. These were linked to the EMIS data using CRRIS (available in the child health information records and the child education records). Around half of the children were linked (47%), but linkage differed by setting, with the highest linkage in Gaza (77%), Syria (72%), and Lebanon (64%), and the lowest in West Bank (33%) and Jordan (31%). When we looked at linkage among those using UNRWA schools (education service users as denominator instead of among UNRWA health access users), coverage increased overall to 90%, and in Gaza (94%), Lebanon (94%), West Bank (87%), Jordan (83%), and Syria (72%).

Reasons for failure to link

Early mortality (as recorded in the obstetric records), and migration soon after birth were hypothesized as the main structural reasons why the obstetric and child health datasets might not link. Unlinked data had a higher percentage of early mortality and an increased presence of children vulnerable to mortality risks such as low birthweight or preterm infants even if their deaths weren't explicitly recorded (Appendix Table 1). Unlinked data also contained a larger proportion of non-refugee mothers as compared to refugee mother. Unfortunately, we couldn't assess migration-related non-linkage.

Over time, unlinked data decreased, likely due to improved reporting, recording, and data entry (Figure 5 and Figure 6). Minimal data errors were identified in sex (1% error), location (0.05% error) or for recording live births as stillbirths (0.006%

Figure 6: Percentage linked by each of the 11 steps, over time (different year of birth cohorts)



M RRIS	Mother refugee registration ID
C MFN	Child medical file number
F RRIS	Family refugee registration ID

error) (Appendix Table 1). We found 69% of multiple pregnancies were of the same sex.

Figure 8 illustrates a decision tree segregating UNRWA-serviced Stage 2 children into various groups based on linkage levels. Three variables mortality, mother's refugee status, and year of birth divided the data into four risk groups. The graph displays (1) group size (percentage of births the group represents out of all births), the percentage unlinked data in each group, the percentage of unlinked data out of all unlinked data.

Mortality (group1) was the smallest group (1% of births) but had the highest prevalence of unlinked data (78%), followed by group2 without mortality recorded but with non-refugee mothers (4% of births with 44% of unlinked data), followed by group 3 (without morality, with a refugee mother, and with year of birth 2010-2012), which had 25% of births and 21% of unlinked data. No mention of mortality and having a refugee mother and a year of birth from 2013 onwards (group4) was the largest group (70% of births) and had the lowest prevalence of unlinked data (12%).

We also quantified the association between a failure to link and variables linked with structural lack of linkage (setting, mother's non-refugee status, recorded mortality, risk factors for early mortality) and to reporting errors (setting, year) using logistic regression models (Appendix Table 2). Syria had the highest odds of data not linking followed by Jordan, West Bank and Lebanon as compared to Gaza. As compared to low risk of mortality group (normal birthweight, term and singleton and not recorded as dead), recorded mortality, and a composite of low birthweight, preterm, or multiples without recorded mortality (as a measure of being a high risk of mortality that may not be recorded) had higher odds of not linking. Non-refugee mothers compared to refugee mother also had higher odds of not linking. The odds of not linking decreased over time.

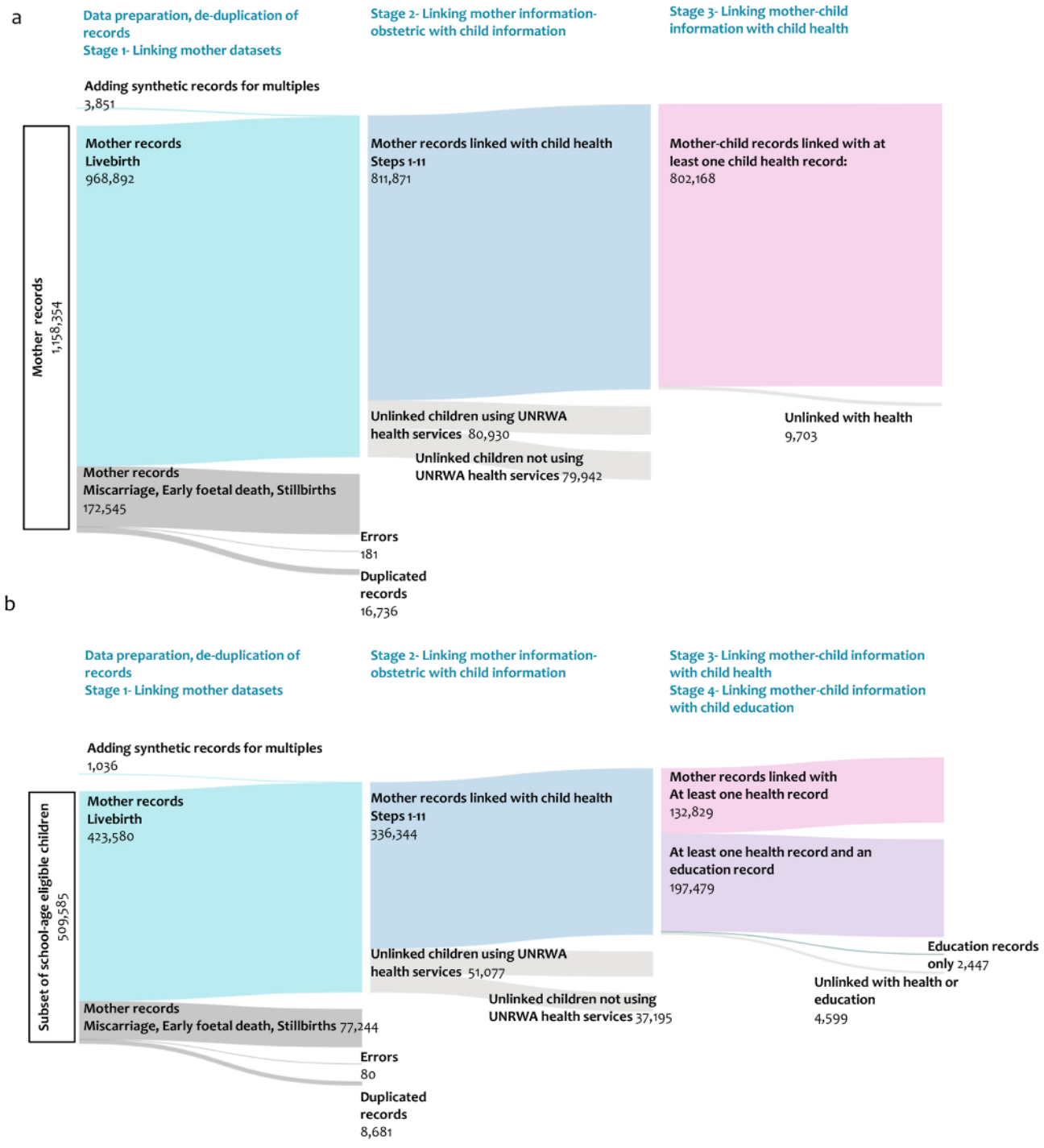
Discussion

We established a birth cohort of Palestinian refugees living in five settings from 2010-2020, using electronic medical records of 972,743 live births, and by linking mother and child health and education records. We found (1) high levels of linkage overall, which improved over time, (2) variations in linkage rates in the five different settings, and (3) factors associated with failure to link including the birth year, setting, mortality record (or risk factors for early mortality) and having a non-refugee mother.

Establishment of a palestinian refugee cohort

Endresen and Øversen (1994) [18] and Zureik and Tamari (2001) [19] have previously noted the research potential of UNRWA's administrative data. Our study is the first use of these data to build a birth-cohort of Palestinian refugees. It provides a significant resource

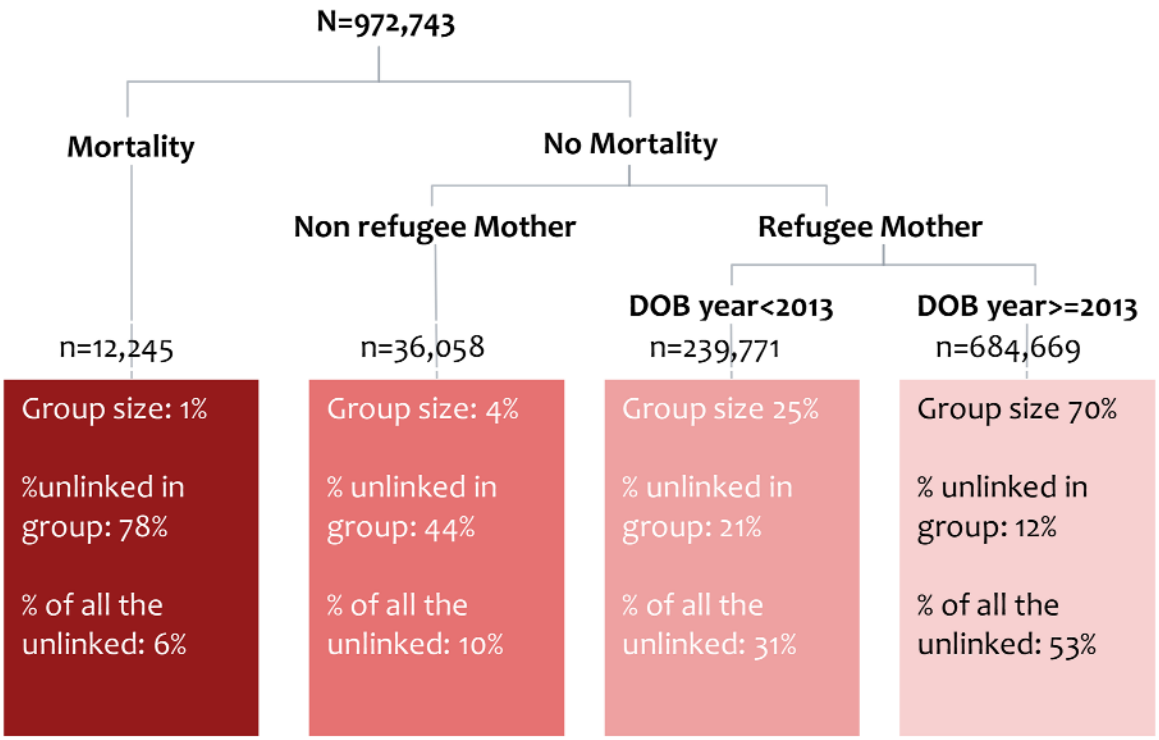
Figure 7: Linkage of (a) maternal records to child outpatient records, (b) maternal records of school-aged eligible children to child outpatient and education records



for future understanding of associations, mechanisms, and problems for protracted refugees and urban poor, filling important gaps in the literature. Victora and Barros note that except for Brazil and India, the top 20 countries publishing on cohorts are all high-income [1]. Since exposures, disease patterns, policies, and health systems differ by setting, our longitudinal dataset will provide new possibilities to study a wide spectrum of policy-relevant questions that apply to urban-poor populations. For example, a 2023 review found that most studies examining the effect of size at birth on subsequent child wellbeing outcomes have been in high-income countries [20] or have not considered size for gestational age; such analyses are possible in our birth cohort. Another unusual feature of our cohort is that it includes five settings and services clustered in 140 health facilities and 702 schools, allowing for context-specific and comparative questions.

Using multi-step deterministic algorithms, we reached a linkage rate of 83% overall for health records, with rates improving from 71% in 2010 to 86% in 2020. This is comparable to other studies linking mothers and children using deterministic methods, for example a linkage rate of 82% in Brazil [21]. The linkage percentage is even higher

Figure 8: CART decision tree to determine the unlinked data



DOB: Date of birth

when defining the cohort as children who use UNRWA health services (91%). We note that linkage improved over time as experience with E-health increased and mis-recording in date of birth decreased (contributions of steps 4/5 and 6/7 to overall linkage decreased). Mis-classification data-entry errors were low (1% error for sex, 0.006% error for the delivery outcome, and 0.05% error for setting). Among children eligible for school, 47% linked, as not all children who used UNRWA health services also went to UNRWA schools. Among children attending UNRWA schools, 90% linked with E-health, indicating that children attending UNRWA schools were more likely to use UNRWA health services. It is possible to explore ways to increase the linkage with education data by loosening the criteria used for linkage (as was done for the mother-child linkage), for example if the date of birth criteria was loosened, as was done for health records in Stage 4.

Characteristics of the population linking

It is essential to recognise that the linked cohort is mainly of those children who used UNRWA services at least once. Access to, and use of, non-UNRWA services differs by setting and is reflected in the percentage of data linking to health and education. More children from Jordan and the West Bank are unlinked (probably because there are alternative choices available for refugee children) while those in Lebanon, Gaza and Syria have fewer options to use non-UNRWA services. In 2019, the Multiple Indicator Cluster Survey in Palestine found that 72% of children aged 5–17 in Gaza accessed UNRWA services, compared to only 22% in the West Bank [22], though this partly reflects the proportions of these populations that are refugees (67% and 30% respectively, see Table 1).

The setting also reflects the timing of the introduction of E-health and the overall quality of record keeping and data entry which can in turn affect linkage. There are several indications from previous work [23] that records from Syria have the poorest recording of birth dates, and that data quality (assessed via digit preference and heaping) are weakest in Syria and Jordan. The conflict situation in Syria has almost certainly impacted the accuracy of the data collected and the linkage process. Migration might prevent the use of children's health services. No studies of numbers of Palestinian-refugee specific migration were found in the literature, but news reports document that the adverse impacts of the conflict in Syria and the economic collapse in Lebanon on Palestinian refugees, have led to drownings during attempted illegal migrations [22, 23].

Linkage improvements over time are most likely to be because the E-Health system improved but may also be due to increased use of free UNRWA services as economic hardship foreclosed other options. Setting thus becomes a complex construct that encompasses both structural conditions, mortality, out-migration and data errors as reasons for non-linkage and for exclusion from the cohort.

Future analysis and recommendations

The established, high-quality birth cohort presents a unique opportunity to explore key research questions concerning Palestinian refugees and urban-poor populations. For future analyses, we point out some considerations to enhance validity.

First, data quality and linkage improved over time, especially from 2013 on, and again from 2017 on, (these years were identified by the CART analyses and were also when UNRWA updated its E-health system). Researchers may wish

to restrict their analyses to data from these years or consider running sensitivity analyses to ensure data quality in early years is not affecting results.

Second, the characteristics of the data that linked need to be understood to avoid selection bias, and properly translate research to policy changes for specific populations. The service-use context suggests that our cohort (of UNRWA service users) is most likely to be generalisable to the entire population of Palestinian refugees in in Gaza, Lebanon and Syria. By contrast, refugees from Jordan and West Bank appeared to use a greater variety of non-UNRWA services or a mix of UNRWA and non-UNRWA services, potentially leading to only the most vulnerable refugees accessing UNRWA services. Our CART analysis also showed children with refugee fathers, but non-refugee mothers, were also less likely to link, possibly because non-refugee mothers could provide their children with access to alternative services. The CART analysis proved to be a useful method for identifying distinct groups within non-linkage data by utilising a combination of different variables and could be used for other linkage studies.

Third, most cohorts in the literature are based in a single country, whereas ours is in five settings (4 countries). This allows for the possibility to examine variations across populations and clusters from five settings, 140 health clinics, and 702 schools. Setting may well be an effect-modifier though, so analyses combining more than one setting need to consider this.

Fourth, multiples (twins, triplets, etc.,) are a challenge in datasets, and many researchers exclude them, even though they are at high risk of adverse outcomes. We retained this important subgroup in our cohort. However, researchers may need to exclude multiples from analyses when using birthweight as an exposure, or alternatively to use imputation methods or sensitivity analyses. This is because some multiples did not have a record for each child, so we created (and flagged) a synthetic record for the second or third neonate (twin or triplet, etc.,) based on the original obstetric record, affecting mainly the birthweight variable where using the same weight for each birth could lead to misclassification. In case of duplicated records with contradicting birthweight results, we propose that future studies using this cohort to conduct a sensitivity analysis on the effect of choosing a different record. Moreover, even when birthweights of all multiples were recorded, we could not be certain which child they belonged to unless the sex was discordant, since the obstetric records had no child ID. Other variables affected by the lack of a child ID (i.e., gestational age, mode of delivery, date of birth, place of delivery) are not problematic because they can be assumed to be identical or very similar for all babies within multiple pregnancies. Moreover, even the potential discordance of birthweight in multiples can be quantified; the literature reports only 16% of multiples have birthweights that are more than 20% different.

Fifth, and finally, this analysis provides opportunities to improve the E-health system. As we showed, examining mortality using our data would require further work to identify deaths and to assess the survival status of children lost-to-follow-up. Child health records do contain a variable to record the date of death, but most deaths occur early, before most neonates are brought into the primary care facilities for services. Deaths are recorded on the obstetric record, but in a free text format that needs cleaning, UNRWA also has a death registration system, but this is voluntary, and families may have little reason to report deaths, leading to under-reporting. This analysis pinpointed to UNRWA the need for a more accurate system to capture mortality data. The multivariable analysis characterising the failure to link found that neonates at higher risk of morality (low birthweight, preterm, or multiple pregnancy) had higher odds of not linking even though they were not reported as dead. This suggests deaths were missed and that researchers interested in mortality will need to examine the full birth cohort (including unlinked data) and include other sources (RRIS) or verification of the survival or migration status of children lost to follow-up.

This dataset offers a tremendous resource for answering important research questions on human capital development of urban-poor and of refugees. Some examples of planned research include exploring the effect of being post-term on size-at-birth and mortality outcomes, the effects of size-at-birth on child obesity, the association between recurrent infection and school performance, or the effects of exposure to conflict or high temperature on birth outcomes and child health and education attainment. However, this study has some limitations. We limited our evaluations of internal data quality to the characteristics of individuals that did or did not link, and to data recording and data-entry error rates based on date of birth/delivery, sex, setting and pregnancy outcome. There was no gold standard to evaluate the true or false matches, or the sensitivity and specificity of the linkage...[24]. UNRWA is consistently seeking to improve its system and this analysis allowed us to pinpoint points of changes to improve the data captured by E-health system.

In future, we hope it will be possible to identify funds to allow data to be shared based on specific requests, subject to review by an independent research review board. Due to the vulnerable position of refugees, and the sensitive nature of their information, utmost care is needed to protect their privacy and to ensure research does not stigmatise them. UNRWA, being the primary collector of personal data of Palestinian refugees, has established a robust data protection system. We would seek to ensure specific components, such as de-identified participant data, linkage procedures, and the statistical analysis plan may be shared for designated analyses to be permitted under a formal access agreement.

Conclusion

We established a Palestinian refugee birth cohort from 2010-2020 using electronic medical records of 972,786 live births, linking mother and child health from 140 primary clinic and education records from 702 schools. We also established criteria for selecting different sub-sets of the cohort depending on the research question and the analytic purposes. Since exposures, disease patterns, policies, and health systems differ by setting, this creates an invaluable resource for future research aiming to elucidate pathways for improved health and education in this vulnerable and understudied population.

Acknowledgement

We acknowledge the UNRWA clinic doctors, midwives, nurses, staff clerks, UNRWA school directors, teachers, and staff who entered the data used to build the cohort. UNRWA Information Technology staff and nurses in Jordan and Lebanon explained the system and the Lebanon field office shared the E-health support booklet used to make the data dictionary. We acknowledge UNRWA staff who supported in the data extraction Mohammad Shraim, Mohammad Habeeb, Anas Alhroub, Nema El-Faleet and Fuad Jadallah. We also acknowledge the SILA research group Sawsan Abdul Rahim, Bassam Abou Hamad, Chaza Akik, Imad El Hajj, Weeam Hammoudeh, Stephen McCall, Nisreen Salti, Aline Semaan, Sam Rose. We also thank Nagasaki University staff who helped administer the WISE grant.

Funding statement

This research was supported by the Nagasaki University "Doctoral Program for World-leading Innovative and Smart Education" for Global Health, KYOIKU KENKYU SHIEN KEIHI, Ministry of Education, Culture, Sports, Science and Technology (MEXT).

Role of funder/sponsor statement

The funder had no role in study design, data collection, data analysis, data interpretation, or writing.

Statement on conflicts of interests

AS, GB, HA, SA, GP, and RI are employed by UNRWA. The other authors (ZJ, MS, HG, OC) declare no competing interests.

Contributors statement

OC conceived the use of UNRWA electronic records for linkage; ZJ, HG, AS, and OC designed the study approach; AS, GB, HA, SA, GP and RI guided the understanding of the dataset structure, GB, HA, SA and RI supported in the extraction of the data; GP encrypted the data; MS, HG, and OC supervised the data analysis; ZJ and OC analysed the data. ZJ wrote the first draft. All authors contributed to the writing of the paper. All authors read and approved the final version.

Ethics statement

Approval to use de-identified encrypted data was obtained from ethics committees of the London School of Hygiene and Tropical Medicine, Nagasaki University, and UNRWA's research review board. No identifiers (apart from encrypted IDs) were shared with researchers outside UNRWA.

References

1. Victora CG, Barros FC. Cohorts in low-and middle-income countries: from still photographs to full-length movies. Journal of Adolescent Health. 2012;51(6):S3-S4. https://doi.org/10.1016/j.jadohealth.2012.09.003

2. Victora CG, Adair L, Fall C, Hallal PC, Martorell R, Richter L, et al. Maternal and child undernutrition: consequences for adult health and human capital. The lancet. 2008;371(9609):340-57. https://doi.org/10.1016/S0140-6736(07)61692-4

3. Victora CG, Christian P, Vidaletti LP, Gatica-Domínguez G, Menon P, Black RE. Revisiting maternal and child undernutrition in low-income and middle-income countries: variable progress towards an unfinished agenda. The Lancet. 2021;397(10282):1388-99. https://doi.org/10.1016/S0140-6736(21)00394-9

4. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584(7821):430-6. https://doi.org/10.1038/s41586-020-2521-4

5. Victora CG, Castro MC, Gurzenda S, Medeiros AC, França GV, Barros AJ. Estimating the early impact of vaccination against COVID-19 on deaths among elderly people in Brazil: Analyses of routinely-collected data on vaccine coverage and mortality. EClinicalMedicine. 2021;38:101036. https://doi.org/10.1016/j.eclinm.2021.101036

6. Christen P, Schnell R. Common Misconceptions about Population Data. arXiv preprint arXiv:211210912. 2021. https://doi.org/10.48550/arXiv.2112.10912

7. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big data & society. 2017;4(2):2053951717745678. https://doi.org/10.1177/2053951717745678

8. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020;47(2):218–26. https://doi.org/10.1080/03014460.2020.1742379

9. United Nations High Commissioner for Refugee (UNHCR). United Nations High Commissioner for Refugee: Figures at a Glance 2022 [Available from: https://www.unhcr.org/figures-at-a-glance.html

10. United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA). Annual operational report 2021. 2021.

11. United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA). UNRWA- Where we work [Available from: https://www.unrwa.org/where-we-work

12. United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA). UNRWA-Departement of health annual report 2021. 2021.

13. United Nations Relief and Works Agency for Palestine Refugees in the Near East. UNRWA Annual Operational Report 2018. 2018.

14. Ballout G, Al-Shorbaji N, Abu-Kishk N, Turki Y, Zeidan W, Seita A. UNRWA's innovative E-Health for 5 million Palestine refugees in the Near East. BMJ Innovations. 2018:bmjinnov-2017-000262.

15. World Bank. Population total 2021 [Available from: https://data.worldbank.org/indicator/SP.POP.TOTL

16. United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA). UNRWA-Departement of health annual report 2022. 2022.

17. Wolfson J, Venkatasubramaniam A. Branching out: use of decision trees in epidemiology. Current Epidemiology Reports. 2018;5(3):221–9. https://doi.org/10.1007/s40471-018-0163-y

18. Endresen LC, Øvensen G. The Potential of UNRWA Data for Research on Palestinian Refugees: A Study of UNRWA Administrative Data.; 1994.

19. Zureik E, Tamari S. Reinterpreting the Historical Records : The Uses of Palestinian Refugee Archives for Social Science Research and Policy Analysis2001.

20. Jamaluddine Z, Sharara E, Helou V, El Rashidi N, Safadi G, El-Helou N, et al. An umbrella review on the effects of size at birth on health, growth and developmental outcomes. Archieves of Disease in Childhood 2022. http://dx.doi.org/10.1136/archdischild-2022-324884

21. Almeida D, Gorender D, Ichihara MY, Sena S, Menezes L, Barbosa GC, et al. Examining the quality of record linkage process using nationwide Brazilian administrative databases to build a large birth cohort. BMC medical informatics and decision making. 2020;20(1):1-9. https://doi.org/10.1186/s12911-020-01192-0

22. Statistics PCBo. Palestinian Multiple Indicator Cluster Survey 2019-2020, Survey Findings Report, Ramallah, Palestine.; 2021.

23. Jamaluddine Z, Paolucci G, Ballout G, Al-Fudoli H, Day LT, Seita A, et al. Classifying caesarean section to understand rising rates among Palestinian refugees: results from 290,047 electronic medical records across five settings. BMC Pregnancy and Childbirth. 2022;22(1):1-13. https://doi.org/10.1186/s12884-022-05264-z

24. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International journal of epidemiology. 2017;46(5):1699-710. https://doi.org/10.1093/ije/dyx177

Abbreviations

ANC:	Antenatal care
C:	Child
E-health:	Electronic Health records system
EMIS:	Education Management Information system
F:	Family
M:	Mother
MFN:	Medical file number
RRIS:	Refugee registration number
UNRWA:	United Nations Relief and Works Agency for Palestine Refugees in the Near East

Appendix Table 1: Hypotheses

Hypotheses			Link	Unlinked
Structural (legitimate) reasons for data not to link				
Early mortality (before the child used UNRWA services or got an MFN)				
1	More deaths among the unlinked	Mortality	0.3%	6.0%
2	More multiples among the unlinked because they have higher early mortality (even if one or both are not registered as a death)	Multiple	1.4%	4.4%
3	More LBW/PT among the unlinked because LBW/PT have higher early mortality (even if not registered as a death)	Preterm (Gestational age <37 weeks)	7.7%	10.9%
		Low Birthweight <2500)	5.6%	9.3%
Child used other services (and never used UNRWA services)				
4	More mothers who are not Palestinian among the unlinked because non-refugee mothers have alternative options for child health and education	Mother RRIS missing in health	2.5%	10.3%
5	More children with a missing MFN (in the mother dataset) are unlinked because children did not use UNRWA services, so a C MFN was not generated	Missing C MFN	0.0%	49.7%
6.a	More families from Jordan and West Bank are unlinked mother and child (because they have more choices). Lebanon, Gaza, and Syria have fewer choices for other services *6ca and (6a or 6b) for in opposite directions*	% Linkage health		
		Jordan	74.8%	25.2%
		Lebanon	89.1%	10.9%
		Syria	67.8%	32.2%
		West Bank	72.8%	27.2%
		Gaza	93.9%	6.1%
6b	More children in Jordan, West Bank do not link to education services because they have more alternative options. Lebanon, Gaza and Syria have fewer choices for other education services	% Link education		
		Jordan	31.9%	68.1%
		Lebanon	63.5%	36.5%
		Syria	72.2%	27.8%
		West Bank	32.6%	67.4%
		Gaza	77.1%	22.9%
Migration (before the child used UNRWA services or got an MFN)				
6.c.	More families from Lebanon and Syria are unlinked (because they have higher migration). Cannot test but might contribute to a higher proportion of unlinked. *Cannot be distinguished from other causes in 4.a.*			

Continued

Appendix Table 1: Continued

Lack of linkage due to reporting, recording or data entry errors			
Data entry errors in any of the IDs (namely Mother/ C MFN, FRRIS.....)			
7	Linkage will improve over time as experience with electronic medical records improved	Figure 5 and Figure 6-Improvement of linkage	
8	Very recent data has more zero in CRRIS as it takes more time to register them	%Missing C RRIS	
		2010	2.0%
		2012	3.8%
		2014	5.7%
		2016	8.8%
		2018	12.0%
		2020	38.9%
Ages mis-recorded/ recorded approximately (heaped on 1 or 15^{th} or in January)			
9	Linkage based on steps +/- 90 and +/- 180 will decrease	Figure 5- decrease in linkage errors	
7	Linkage will improve over time as expertise in electronic medical records improved	Figure 5 and Figure 6- Improvement of linkage	
Sex mis-recorded			
10	Attempt to link unlinked kids to any sex.	A total of 13,683.	Error 1.4 %
Location mis-recorded			
11	Attempt to link unlinked kids to any location	A total of 477 links. Error 0.05%	
Live birth miscoded as stillbirth (so was excluded from the start)			
12	Attempt to link unlinked children to stillbirths	A total of 56 links. Error 0.006%	
Stillbirth miscoded as a live birth			
	Cannot test	Might be like step 12	
Distinguishing of duplicated records from multiples			
13	The percentage of same-sex multiples.	The sex ratio observed in the data is 1.03 male (50.7%) to 1 female (49.3%). In our dataset same sex multiples (69%); discordant multiples are 31%. This 69% is plausible if we assume that ∼30% of multiples are monozygotic (so same sex as per published reports) and around half of dizygotic multiples are same sex $(0.3+(0.7(0.5068^2+0.4932^2)) = 0.30+0.35 = 65.0\%$ of multiples expected to be same sex.	

Appendix Table 2: Multivariable logistic regression model of the association of different population characteristics of children using UNRWA services and odds of linkage (N = 892,801)

		Adjusted OR (95%CI)
Setting	Gaza (ref)	1.0
	Jordan	3.2 (3.1-3.2)
	Lebanon	1.4 (1.3-1.4)
	Syria	4.3 (4.2-4.4)
	West Bank	2.4 (2.3-2.4)
Mother ID	Refugee (ref)	1.0
	Not a refugee	2.7 (2.6-2.8)
Dead or at risk of early mortality	Normal birth weight, term and singleton and not recorded as dead (ref)	1.0
	Low birthweight, or preterm or multiple, and not recorded as dead (at risk of early mortality)	1.6 (1.6-1.6)
	Recorded as dead	47.0 (44.8-49.3)
Year of birth	2010	1.0
	2011	0.8 (0.8-0.8)
	2012	0.6 (0.5-0.6)
	2013	0.4 (0.4-0.4)
	2014	0.4 (0.4-0.4)
	2015	0.3 (0.3-0.3)
	2016	0.3 (0.3-0.3)
	2017	0.2 (0.2-0.2)
	2018	0.2 (0.2-0.2)
	2o19	0.2 (0.2-0.2)
	2020	0.2 (0.2-0.2)