Get every one in the picture

Data Quality 2

Workshop on data analysis and report writing for civil registration based vital statistics

Nadi, Fiji
30 January – 03 February 2023

Bloomberg Philanthropies | DATA FOR HEALTH INITIATIVE

Pacific Community
Communauté du Pacifique

Australian Bureau of Statistics

UNITED NATIONS ESCAP
Economic and Social Commission for Asia and the Pacific

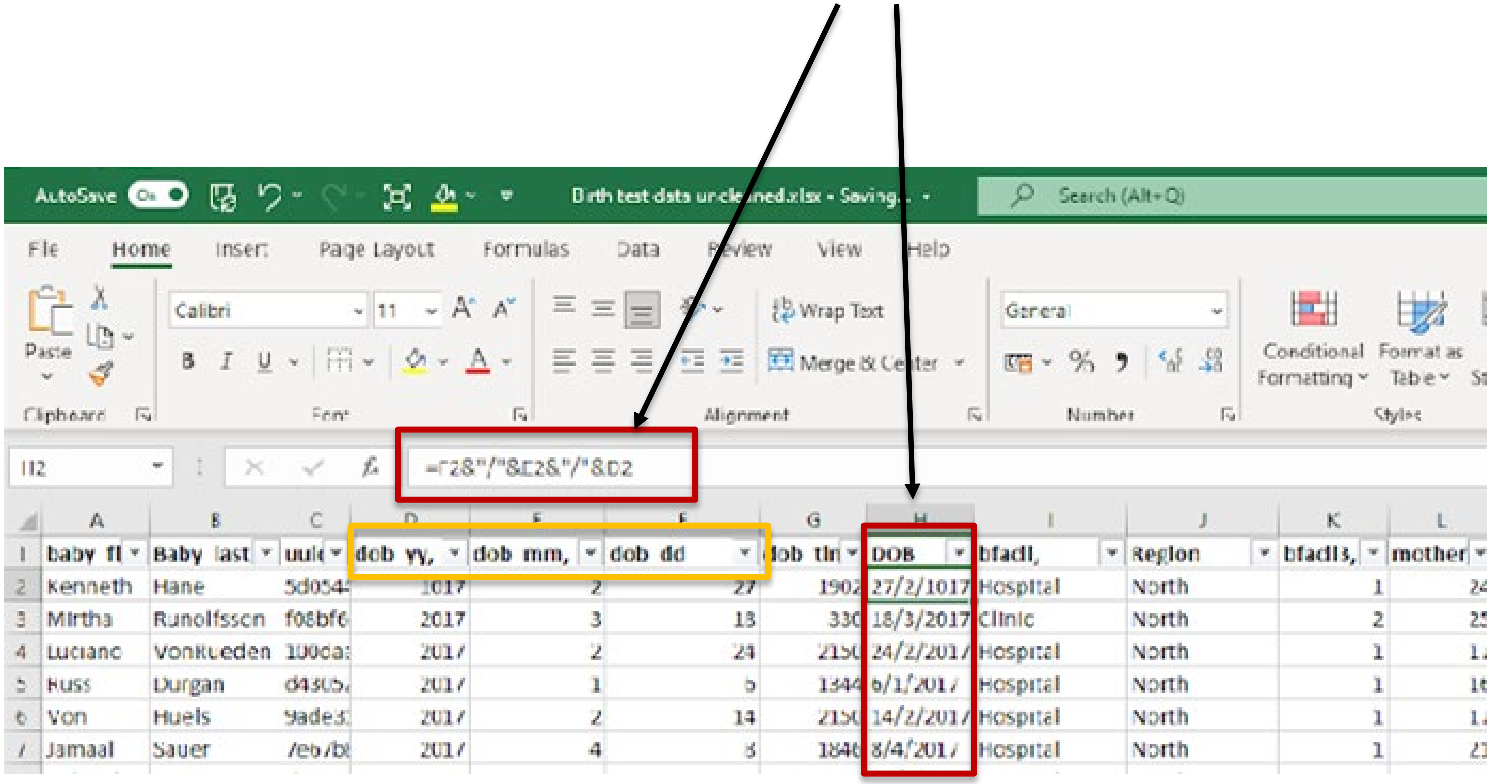
Data Quality Checks

- Concatenating/splitting variables
- Creating categorical (grouped) variables
- Checking for implausible values
- Re-distribution for missing values for age at death or mothers age at birth
- Aggregates and multi-year estimates
- Off-island events

Concatenating variables

Can be used to combine information in multiple variables, into one variable.

E.g. If date of birth in your dataset is separated by day, month and year, you can use the following formula in excel to combine them to create one DOB



I12

=F2&"/"&E2&"/"&D2

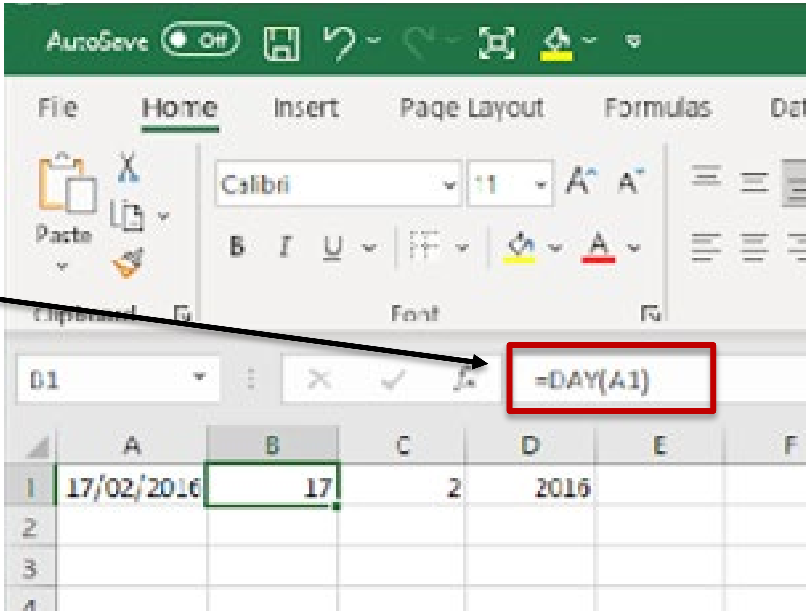
	A	B	C	D	E	F	G	H	I	J	K	L
1	baby fi	Baby last	uuid	dob yy,	dob mm,	dob dd	dob tin	DOB	bfacil,	Region	bfacils,	mother
2	Kenneth	Hane	5d0542	2017	2	27	1902	27/2/2017	Hospital	North	1	24
3	Mirtha	Runolfsson	f08bf6	2017	3	13	330	13/3/2017	Clinic	North	2	25
4	Luciano	VonRueden	100ca3	2017	2	24	2150	24/2/2017	Hospital	North	1	1
5	Russ	Durgan	d43c5	2017	1	6	1344	6/1/2017	Hospital	North	1	16
6	Von	Huels	9ade3	2017	2	14	2150	14/2/2017	Hospital	North	1	1
7	Jamaal	Sauer	7eb7b	2017	4	8	1846	8/4/2017	Hospital	North	1	21

Splitting a variable (DOB)

If date of birth is formatted as a single date in your dataset, you will need to split this variable into three separate variables: DAY/MONTH/YEAR

Note: Ensure that your date variable is formatted as a date in Excel

In a blank cell, insert the formula =DAY(CELL) to compute the day from the date of birth. You can then repeat the process for month, using the formula =MONTH(CELL) and for year, using the formula =YEAR(CELL)



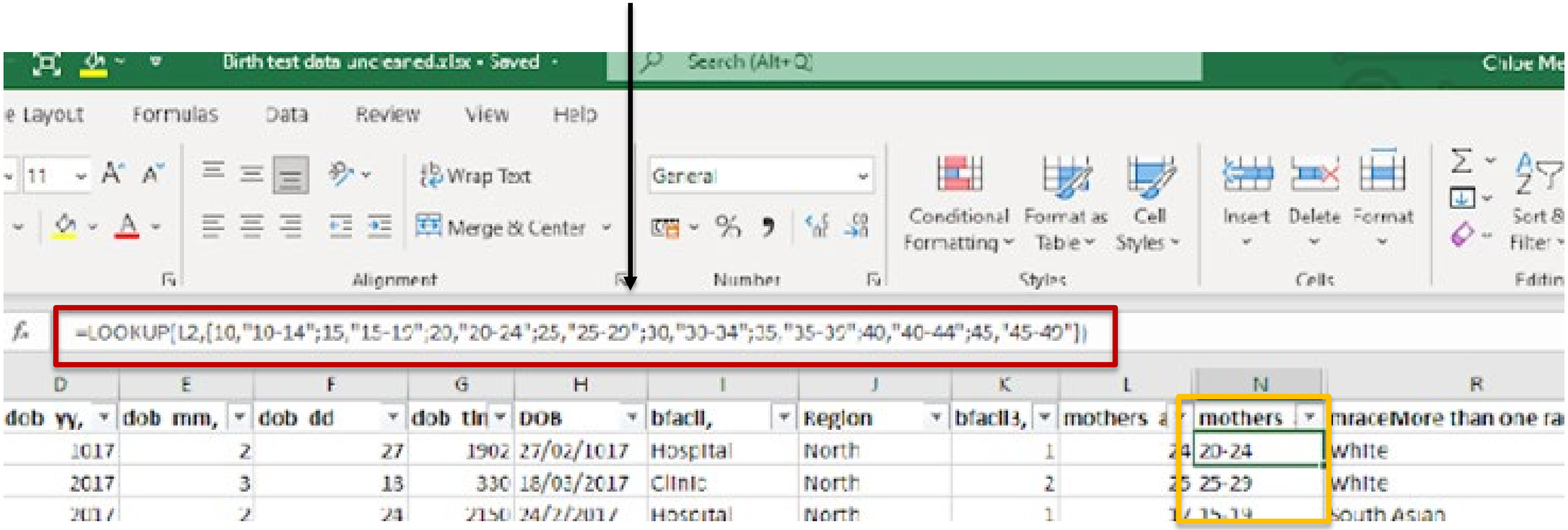
=DAY(A1)

	A	B	C	D	E	F
1	17/02/2016	17	2	2016		
2						
3						

Creating categorical (grouped) variables

If you have numerical data that you would like to group, you can use the ‘LOOKUP’ function in Excel.

E.g. This may be useful for grouping age. The example we are using here is for creating a grouped variable for mother’s age at birth. Note how the formula takes the information on mother’s age from Cell L2.



```
=LOOKUP(L2,{10,"10-14";15,"15-19";20,"20-24";25,"25-29";30,"30-34";35,"35-39";40,"40-44";45,"45-49"})
```

D	E	F	G	H	I	J	K	L	N	R
dob yy,	dob mm,	dob dd	dob tin	DOB	bfacil,	Region	bfacil3,	mothers a	mothers	mraceMore than one ra
2017	2	27	1902	27/02/2017	Hospital	North	1	24	20-24	White
2017	3	18	330	18/03/2017	Clinic	North	2	25	25-29	White
2017	2	24	2150	24/2/2017	Hospital	North	1	17	15-19	South Asian

Implausible values

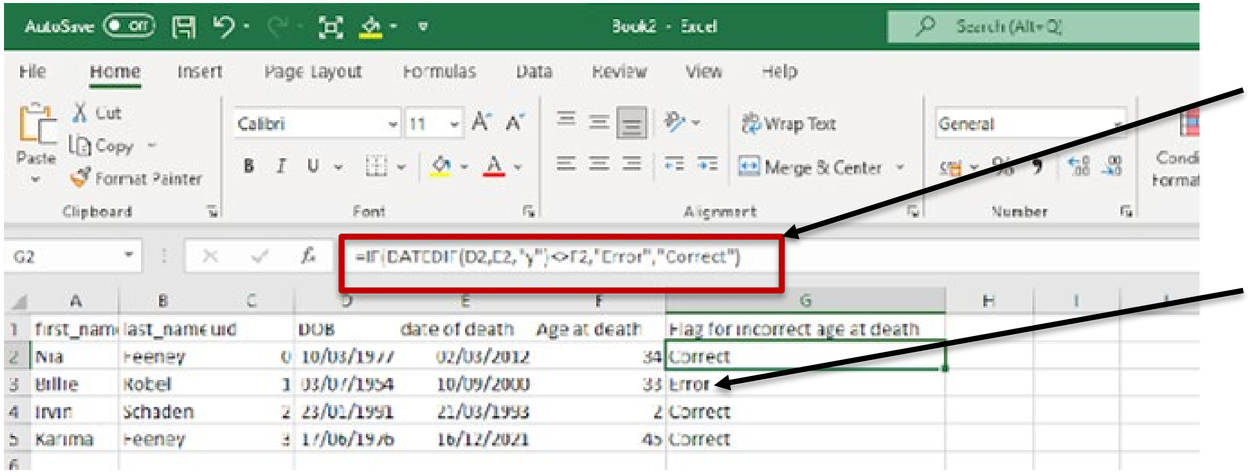
- Are there any odd patterns of births by age of the mother?
  - E.g. Births to mothers aged younger than 10 years or over 49 years may warrant further investigation

- Are there any out-of-range values for date of birth and age of mother/age of deceased?
  - E.g. **Age at death** does not match with **Date of Birth** and **Date of Death**

Checking for implausible values

We can mostly use pivot tables and/or filter functions to check for implausible values, for example if we want to see if there are any implausible values for mothers age at birth.

However, if we want to check the plausibility of some variables against each other, we may need to use Excel formula to create a flag variable which will notify you of any erroneous data. Here is an example of checking for implausible values for age at death when we have the date of birth and date of death for the deceased:



G2

=IF(DATEDIF(D2,E2,"y")<>F2,"Error","Correct")

	A	B	C	D	E	F	G
1	first_nam	last_name	uid	DOB	date of death	Age at death	Flag for incorrect age at death
2	Nia	Feeney	0	10/03/1977	02/03/2012	34	Correct
3	Billie	Kobel	1	03/07/1954	10/09/2000	33	Error
4	Irvin	Schaden	2	23/01/1991	21/03/1993	2	Correct
5	Karima	Feeney	3	17/06/1976	16/12/2021	45	Correct
6							

This formula creates a flag which checks the age at death against the date of birth and date of death. You'll notice here that it has identified an erroneous age at death for the second person in the dataset.

Re-distribution for missing age at death

- When death records are missing age, we need to estimate age at time of death.
  - This also applies to births where age of mother is unknown

- Uses the age distribution of deaths with known ages to determine how many deaths of unknown age should be placed in each age group.

- Age patterns are different for males and females, so the re-distribution of these deaths should be computed separately by sex.

- Whether the re-distribution is done by year, or over an aggregated period will depend on the overall number of deaths, and the proportion for which no age is reported.

Re-distribution for missing age at death

- Start by setting up a table of deaths by age group and sex for the year(s) where data are missing

- For the deaths for which age is known, calculate the percent distribution of these deaths by age group for each sex separately

- Multiply the percent for each age group from this distribution to the total number of deaths (including deaths of unknown age) to get the revised number of deaths by age

- Round your results to the nearest whole person (after all – we don’t get part of a person dying!)

- This method can also be used to re-distribute births by age of mother

- **Refer to sheet T3.11 in the Excel workbook for tables which have been pre-prepared for you to input data to compute redistributed values for mother’s age at birth and age at death**

For the deaths for which age is known, calculate the percent distribution of these deaths by age group for each sex separately

Age	Total deaths		Percentage of total excluding unknown ages (%)		Re-distributed deaths by age	
	M	F	M	F	M	F
<1 year	14	12	3.8	3.9	15	13
1-4	6	4	1.6	1.3	7	4
5-9	2	1	0.5	0.3	2	1
10-14	1	4	0.3	1.3	1	4
15-19	5	6	1.4	2.0	5	6
20-24	9	13	2.5	4.3	10	14
25-29	16	12	4.4	3.9	17	13
30-34	23	12	6.3	3.9	25	13
35-39	25	14	6.8	4.6	27	15
40-44	22	15	6.0	4.9	24	16
45-49	26	22	7.1	7.2	28	24
50-54	35	26	9.6	8.5	38	28
55-59	38	28	10.4	9.2	41	30
60-64	48	32	13.1	10.5	52	35
65-69	58	44	15.8	14.4	63	47
70-74	36	36	9.8	11.8	39	39
75+	2	24	0.5	7.9	2	26
Unknown	33	24				
TOTAL	399	329	100.0	100.0	399	329

Apply this percentage to all deaths (including deaths with unknown age)

$$= \frac{14 \text{ deaths}}{(399-33) \text{ deaths}} \times 100$$

$$= \frac{14 \text{ deaths}}{366 \text{ deaths}} \times 100$$

$$= 3.8\%$$

$$= \frac{3.8 \times 399}{100}$$

=15.16 deaths to males aged <1

Aggregating data over time

- The very small populations of PICTs and subsequently the small numbers of births and deaths can result in poor consistency of data over time due purely to **stochastic or random effects**.

- Single year data for indicators is likely to be unstable and is not recommended due to small population sizes and disaggregation for sex, age groups, and causes.

- Data should be **aggregated** over 3-5 years before calculating measures of fertility or mortality to account for this.

- Depending on your data, determine whether 3 or 5 year aggregation is appropriate.

- It is preferable that the same time periods are used for births and for deaths for consistency.

Data quality of tabulated data

- Once our unit record data is as good as it can be, we can make judgments about the quality and reliability of our tabulated data.

- Identify the most important sources of error and provide quantitative measures where possible or qualitative descriptions otherwise.

- Report specific sources of error or bias to inform readers.

- Make sure to note down any amendments you make to your data as you go along!

Off-island events

- Deaths occurring off-island impact the distribution of causes of death, and subsequently bias results if not considered.
  - (or in the case of Guam, New Caledonia, and possibly Fiji – deaths in patients that have been referred for treatment from other islands)

- For example, cancer patients referred overseas for treatment may die overseas.
  - May not be counted in their home countries records (as a death certificate would be issued in the country where the death occurred)
  - The result is proportional mortality for cancer may be under-represented on their home island.

- When estimating civil registration completeness, it is advised that overseas births and deaths are not included in civil registry figures since the other data sources to which they will be compared (health data), usually do not include overseas events

Get every one in the picture

Q&A