

## Exercise 5: Data entry and validation

At the end of this exercise you should be able to:

- Know the three ways of reducing data entry errors
- Copy the structure of a REC file
- Export data from EpiData files
- Validate duplicate data files

You have a line listing of 15 records on the pages following the task. The data from these should now be entered. But before you start working, a few considerations are in place.

### Ensuring quality data entry

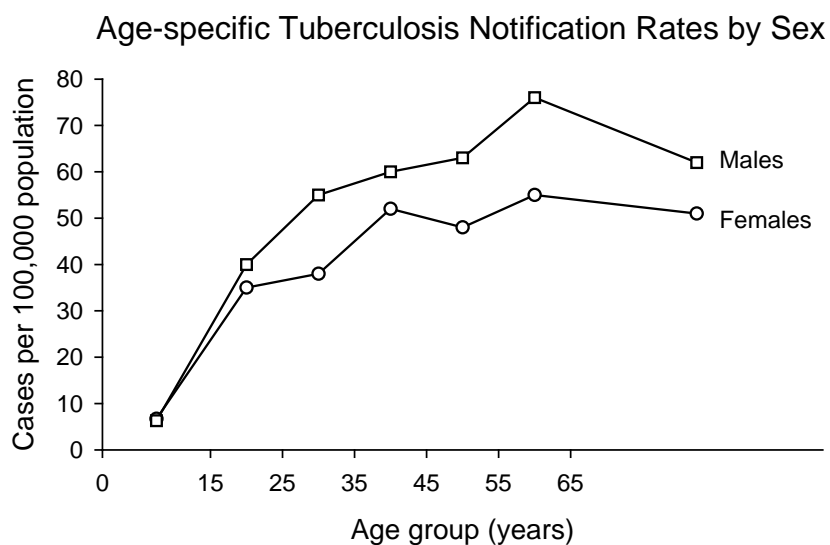
The motto for this course is:

*“You wish never to find yourself in a position to defend the quality of your data”*

Michael B Gregg, formerly MMWR Editor, deceased

You might be challenged about the interpretation of your data, that is part of the scientific process, but your data should be of impeccable quality.

What do you think about the following graph?



It looks nice and we could talk about the differences between males and females and this and that. But we will keep it short: it is total nonsense. The data underlying this graph have no basis, they were made up. Of course, if we were to present these data for real, it would be outright scientific fraud. Few people commit that (but it exists). **Nevertheless, often no assurance can be given that the computerized data are a true reflection of the original data source.** People may have in all honesty done “their best” and assume that they made no errors or so few that it really doesn’t matter. However, this is not good enough for science in general and public health and epidemiology in particular.

There are three ways how we reduce and ultimately eliminate data entry errors:

- o Using a \*.CHK file
- o Working together
- o Duplicate data entry and validation

### Using a \*.CHK file

We have already a few conditions inbuilt that limit data entry errors by creating the A\_EX04.CHK file. For instance, a MUSTENTER field will prevent a data entry person to skip an actually recorded value, as one cannot continue without having entered a value for that field. For the field SEX, we allowed only 1, 2, and 9 as legal values. It is thus not possible to enter “3” into this field. Combined with the pop-up menu during entry, no confusion can arise. The \*.CHK file is an extremely powerful tool to control how data are entered.

### Working together

Entering data alone requires shifting attention constantly between the paper record and the computer screen. This will almost by necessity result in numerous errors, being it that one record is skipped or that it is forgotten what we just read. It should be routine that two persons work on data entry: one person reads aloud the Field value, the other repeats it aloud and enters the value.

### Duplicate data entry and validation

Even with both of the above precautionary measures, data entry errors will still occur, and worse, to an unknown extent. ***The only way, and the only acceptable one, is to enter the data twice into two different files, and then to compare the two files for discordances.*** Any discordance uncovered will then be corrected against the original paper record.

EpiData Entry provides this powerful tool and it offers two approaches to it. The first approach is to enter the data independently twice. The second approach is to prepare for duplicate entry. After the first file is completed, the second file is prepared based on a key field for the first file. While then entering the second duplicate file, the value is checked for each field in each record against the same record of the first file while entering it and you are warned of any discordance, so that you can ensure proper recording during the second entry process.

In either case, we need a unique identifier. We have made a provision that we have such an identifier (see Exercise 4). Sometimes, it must be constructed from more than one variable, an approach you are going to learn later. Laboratory numbers are serial, and it is thus usually seen whether they are unique, but with some other identifiers this is not the case. We will

thus use an EpiData Entry command that will ensure that EpiData checks every “unique identifier” whether it is really unique. To the field for the serial number we thus added the following command in the previous exercise:

```
serno
  MUSTENTER
  KEY UNIQUE 1
END
```

If a duplicate key is revealed, then a data entry note (a \*.not file) should be written. This can be invoked with **F5**. In this note, you would exactly specify with what identifier you have replaced the duplicate key, so that this note can be given to those who enter the data the second time, enabling them to use the same alternative key.

At this point in time, you will be using the first approach, and that is to independently enter the 15 records twice and then to compare the two files.

**Note for data entry:** Do never move around the fields with the help of the mouse. The mouse movements can not be recorded properly and unforeseen errors may occur (e.g., bypassing a calculation made in a field, missing a MUSTENTER command, etc), because the Check file cannot be applied to fields you skip by moving the mouse from one to another. Use only TAB, cursor keys and the Enter key to move around an EpiData entry form.

### **Ensuring that we have a unique identifier before saving the record to disk**

We can do all pleading (above) not to use the mouse or not to save an unfinished record, but data entry persons are bound to perhaps defeat all pleas. It is thus best to do something in the CHK file that makes it failsafe. While one would hope that the mouse is never used to skip a field, if it is done, then one has missing information for the field which is bad. But not having a unique identifier is close to disastrous as one would see when validating data. The following commands at the end of the CHK file will prevent the data entry person to save a record without a unique identifier (SERNO in this case) (see end of previous exercise):

```
AFTER RECORD
  IF serno=. THEN
    HELP "Core information missing:\n SERNO\n must be available" TYPE=WARNING
    GOTO serno
  ENDIF
END
```

### **Tasks:**

- o **Take your A\_EX04.REC file, go to “Tools” “Copy structure” and copy the A\_EX04.REC including its A\_EX04.CHK file to:  
A\_EX05 A.REC and A\_EX05A.CHK files  
A\_EX05 B.REC and A\_EX05B.CHK files**
- o **Enter the 15 records using the A\_EX05A.REC file.**

**After completing data entry, enter the same data again into to the A\_EX05B.REC file.**

- o After you have completed the two files, go to “5. Document” “Validate Duplicate Files” and produce a \*.NOT file giving you a list of any discordance. Save the \*.NOT file as A\_EX05AB.NOT*
- o Use “6. Export Data” “Epidata” to export either one of the two files to a new A\_EX05F.REC file, and then make all corrections in this file. This is your final dataset.*

**On the next page you find the dataset with 15 records**

Laboratory: Ganda Chivua

## Tuberculosis laboratory register

Year: 2002

Lab Serial No.	Date specimen received	Name	Sex M/F	Age	Name of referring facility	Address - patient for diagnosis	Reason for examination*		Results of specimen			Only for SS+ for diagnosis: TB Number or BMU**	Remarks
							Diagnosis (tick)	Month of follow up	1	2	3		
3298	26 Oct	Mary	F	35	Bindura	Beijingstr. 6		5	neg	neg			
3299	26 Oct	John	M	20	Awuna	Tokyo Ave 5	√		neg	neg	neg		
3300	26 Oct	Petra	F	30	Birchenough	Bangkok Rd 108		5	neg	neg			
3301	26 Oct	Charles	M	24	Bindura	Hanoi Street 7a		2	neg	neg			
3302	26 Oct	Tiffany	F	38	Bindura	Hongkong Ave 8	√		neg	neg	neg		
3303	26 Oct	George	M	60	Bindura	Zurich Rd 923	√		neg	neg	neg		
3304	26 Oct	Luke	M	78	Awuna	Paris Street 18a	√		neg	neg	neg		
3304	26 Oct	Virginia	F	28	Birchenough	London Rd 24	√		neg	neg	neg		
3305	27 Oct	David	M	50	Awuna	Baltimore Str 1		6	neg	neg			
3306	27 Oct	Hans	M	50	Ganda Chivua	Bern Str 12	√		1+	1+	1+	Ganda Chivua No 342	
3307	27 Oct	Bill	M	68	Bindura	Berlin Ave 88	√		neg	neg	neg		
3308	27 Oct	Susan	F	29	Birchenough	Amsterdam Rd 3		5	neg	neg			
3309	27 Oct	Marc	M	36	Bindura	Vienna Str 76		2	neg	neg			
3310	27 Oct	Eve	F	15	Awuna	Rome Ave 4		5	neg	neg			
3311	27 Oct	Anthony	M	37	Birchenough	Antwerp Str 26c		6	neg	neg			

\* Check the appropriate category from the *Request for Sputum Examination*

\*\*TB register number or name of the referral BMU (Basic Management Unit)