Exercise 5: Data entry and validation

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

a. Know the three ways of reducing data entry errors
b. Copy the structure of a EPX file
c. Export data from EpiData files
d. Validate duplicate data files

You have a line listing of 15 records on the page following the task description. These data should be entered in this exercise. But before you start working, a few considerations are in place.

Ensuring quality data entry

The motto for this course is:

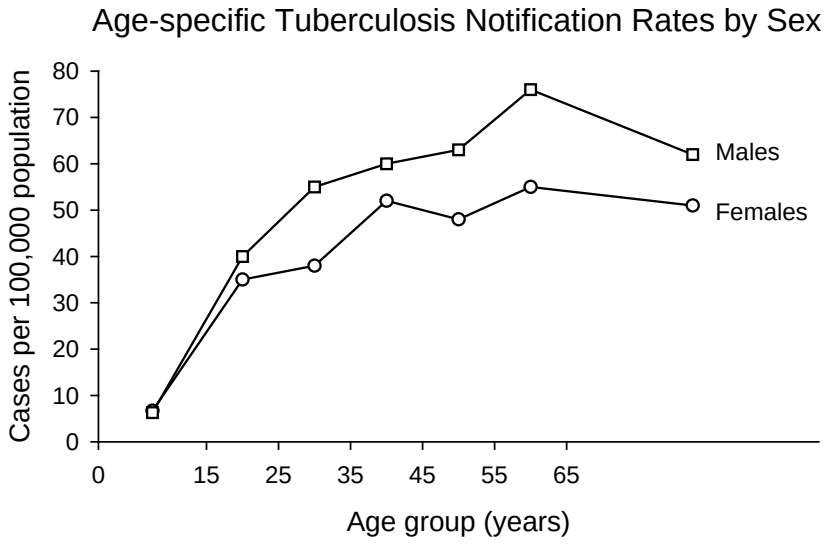
"You wish never to find yourself in a position to defend the quality of your data"

Michael B Gregg, formerly MMWR Editor, deceased

See also: Gregg M B. Field epidemiology (2nd ed). Oxford University Press. 2002: p 414

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 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:

- Defining well thought-through data entry controls in the EpiData Manager
- Working together
- Duplicate data entry and validation

Defining well thought-through data entry controls in the EpiData Manager

We have already a few inbuilt conditions that limit data entry errors by creating the `a_ex04.epx` file. For instance, a `Must Enter` 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 controls set in the EpiData Manager are an extremely powerful tool to control how data entry can be controlled through restrictions.

Working together

Entering data alone requires continuously shifting attention between the paper record and the computer screen. This will almost by necessity result in numerous errors, be it that a 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 the entry with the original paper record.

The rationale behind this process is: *the probability of committing the same error in the same field twice when data entry is done independently by two persons is very small.* Hence, if we list all the discrepancies by comparing the two databases and correct all of them, then we can be reassured that the remaining frequency of data entry errors is miniscule.

EpiData provides this powerful tool and we need a unique identifier to do this. We have made a provision that we have such an identifier (see previous exercises). Sometimes an identifier must be constructed from more than one variable as we have shown.

If a duplicate key is revealed (because there is a perhaps a problem with a component contributor), then a data entry note should be written, best as a text file that is kept open during data entry and amended as one proceeds. In this note, you must specify exactly with what identifier you have replaced the duplicate key, so that this note can be passed on to those who

enter the data the second time, enabling them to use the same alternative key when the necessarily stumble over the same problem.

Before you get to actually enter the data, you find here some assistance to make your data entry work more efficient.

Make duplicate EPX files

As we are entering the same data twice, we need two sets of the `*.epx` files, one for the first, the other for the second entry.

Task:

- ***Download the solution of Exercise 4 and save the file as `a_ex05_a.epx` and `a_ex05_b.epx` files***

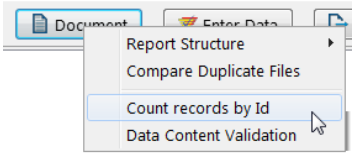
Double-entry

Enter the 15 records into the `a_ex05_a.epx`, then repeat data entry with the `a_ex05_b.epx` file.

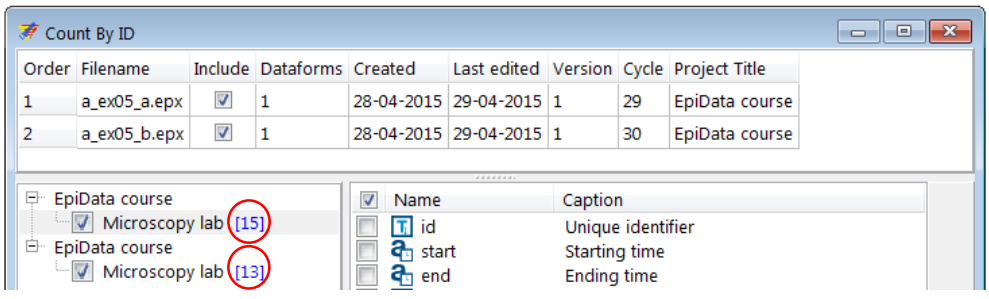
Data validation

After completing double-entry, the two data files are compared, a procedure termed "**data validation**".

The first thing to verify is that we have the same number of records in both sets. If that is not the case, then this must be fixed first. Here for instance, we chose to count the records:

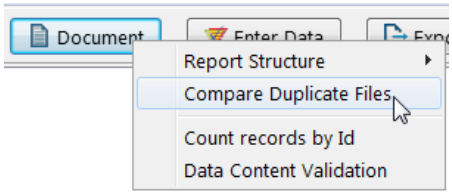


and found them to be unequal:

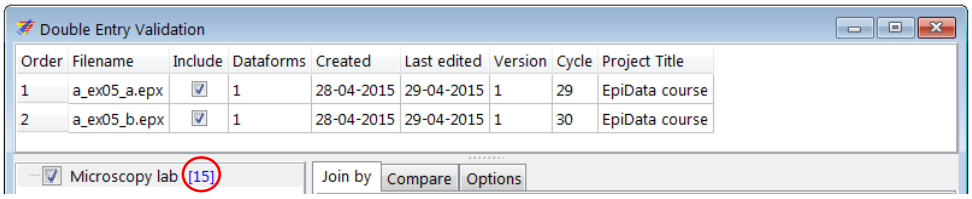


While we know from this small dataset that there actually are 15 records, and thus that the first set has the correct number and the second is missing 2, knowing the expected number of records is the exception in real life. It may also very well be that both sets have the same number of records but both are short of records because by chance each set is missing the same number of

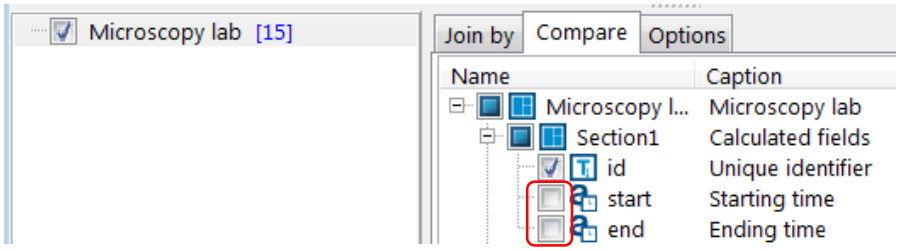
records. We thus have to check which records are missing and we do this by a validation, actually comparing the two sets:



The first of the two files is the "referent" and it is the number of records in this set that is displayed:



In the tab Compare, we untick the computer-provided time fields because they are likey to vary for virtually every record between the two files and would undesirably be listed as discordances:



In the report, we move to the Overview and find the confirmation that the duplicate file (the *_b.epx file) has 2 records missing:

```
---------------------
Result of Validation:
---------------------

Overview
------------------------------------------
Test                                Result

Records missing in main file             0
Records missing in duplicate file        2
Non-unique records in main file          0
Non-unique records in duplicate file     0
Number of fields checked                12
Common records                          13
Records with errors                      1
Field entries with errors                1
Error percentage (#records)           7.69
Error percentage (#fields)            0.64
------------------------------------------
```

Moving further down, we find the serial numbers missing from the second file and thus know which two records must be added first before a proper validation is possible:

```
Record no: 14
Key Fields:
 lab = ML J
 serno = 3310
 regyy = 2003     Record not found
-------------------------------------
Record no: 15
Key Fields:
 lab = ML J
 serno = 3311
 regyy = 2003     Record not found
-------------------------------------
```

The Validation report

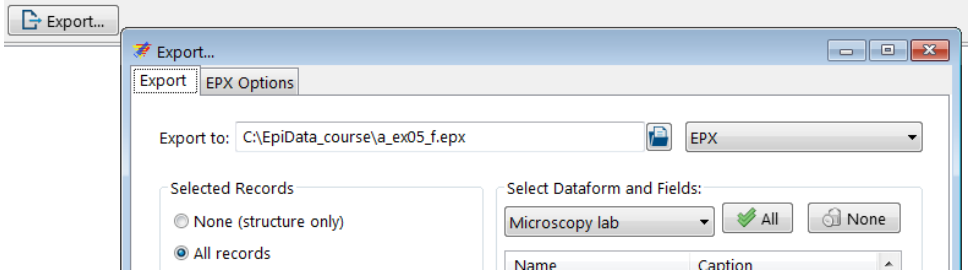
We add these two records and then start again the validation process. This time we get the same number of records in the two files and find one discordance in `serno 3302`:

```
---------------------
Result of Validation:
---------------------

Overview
-------------------------------------------
Test                                Result
-------------------------------------------
Records missing in main file             0
Records missing in duplicate file        0
Non-unique records in main file          0
Non-unique records in duplicate file     0
Number of fields checked                12
Common records                          15
Records with errors                      1
Field entries with errors                1
Error percentage (#records)           6.67
Error percentage (#fields)            0.56
-------------------------------------------

Datasets comparison:
-----------------------------------
Main Dataset:    Duplicate dataset:
-----------------------------------
Record no: 5     Record no: 5
Key Fields:
 lab = ML_J
 serno = 3302
 regyy = 2003
Compared Fields:
sex = 2          sex = 1
-----------------------------------
```

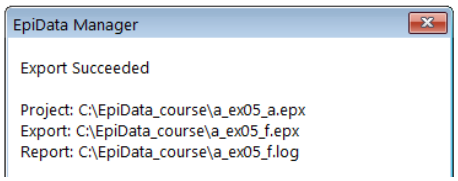
It is essential that this validation report is saved to ensure having a permanent record of the validation process. We propose to save it as a text file `a_ex05_validation.txt`.

Creating a final dataset

One might be tempted to make corrections of any errors that might be identified through discordances in either the `*_a.epx` or the `*_b.epx` file. Doing so would, however, break the "chain of evidence": you could never repeat the validation process and get the same result, but data quality-assurance requires that the validation process is actually exactly reproducible. Therefore, the corrections must be made in a third file. To this end, we export the data from one of the source files to another EpiData file that we will call the `a_ex05_f.epx` file. To standardize as many things as possible, we always export the `a_ex05_a.epx` file to the `a_ex05_f.epx` file (even if in fact it is irrelevant whether we use the `a_ex05_a.epx` or the `a_ex05_b.epx` file, but consistency is good policy). We thus select from Export the `a_ex05_a.epx` file and define in the menu the name and type and All records:



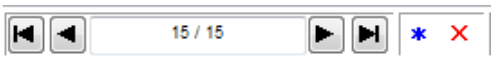
We get a report that export was successful:



Back from Manager to the EntryClient we open `a_ex05_f.epx` and search the record with `serno 3302`.

How to navigate through an `*.epx` file?

To navigate between the records of the `*.epx` file use the navigation panel on the left bottom end of the data entry screen which can be used to navigate through the records.



Shown vertically one by one:

 Go to first record

 Go to previous record

11 / 15 This is record 11 out of a total of 15 records

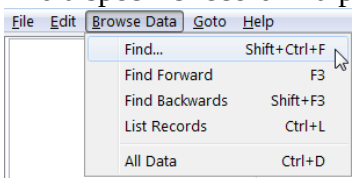
 Go to next record

 Go to last record

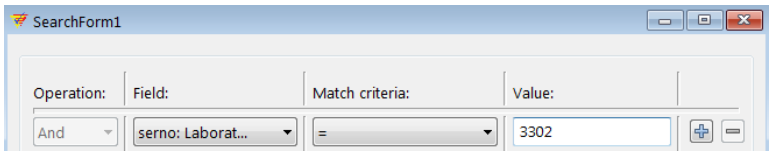
* Insert a new record

× Mark current record for deletion

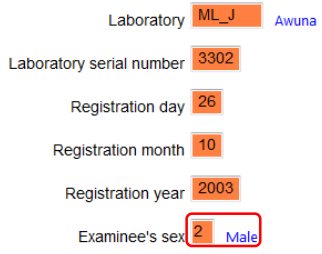
This is useful for a quick forth and back, mainly during data entry, but for the current task to find a specific record in a potentially large file, we use Browse Data | Find:



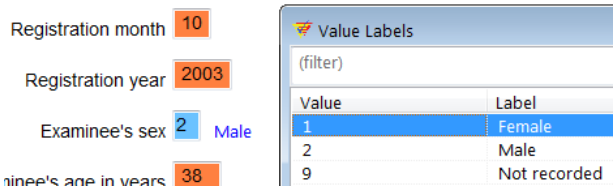
We enter our criterion:



and get thus to the record:



Comparing with the original paper record, we see that the true `Examinee's sex` should be `female`. Moving the cursor into the field and pressing **F9**, we can now pick the correcting value:



As this is the only discordance we save the revised record and exit. We now have a validated file with all discordances resolved.

How to delete a record?

Deleting a record consists of two steps – first, marking a record for deletion; second, permanently deleting it. This is just a safety feature in EpiData to ensure the deletion of record does not happen by chance.

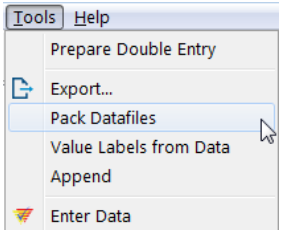
Steps in marking a record for deletion (Look at the screenshot below)

1. Open the `*.epx` file and go to the record you want to delete.
2. Click on the red 'cross' mark next to the navigation panel at the left bottom of the data entry screen. The word `DEL` appears at the side of the red 'cross' mark.
3. Click the arrow in the navigation panel to go to the next record. This will prompt you to save the record. Click '`Yes`' and this successfully marks the record for deletion.
4. Note that the record is not yet permanently deleted from the database. If you realize that this record was not to be deleted, you can undo the action by clicking on the same button and saving the record. `DEL` will disappear now: the red "cross" is a toggle key:

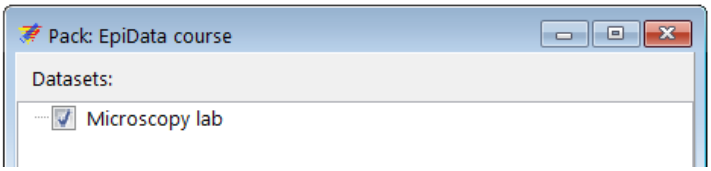
Click on the red cross	DEL appears	Click on the red cross again and DEL disappears
		

How to permanently delete a record? (Pack Data Files)

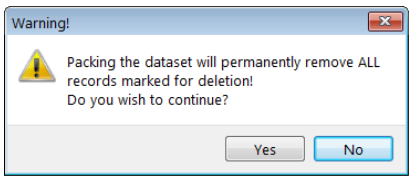
For data protection, it is not foreseen to permanently delete a record in the EpiData EntryClient. This must be done in the EpiData Manager. Close thus the file (if open) in EpiData EntryClient and go to the Manager to Tools | Pack Datafiles:



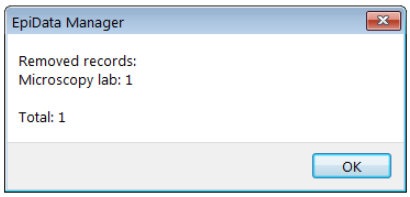
Choose the data file and tick the data form (it could be a relational data base with several forms and you got to choose which one):



Because this is going to be permanent, you receive a final Warning:



After accepting, you get confirmation that it is done:



No backup file is written, this is permanent.

Tasks:

- ***Download the solution of Exercise 4 and save your*** `a_ex04.epx` ***file as*** `a_ex05_a.epx` ***and*** `a_ex05_b.epx`***.***
- ***Enter the 15 records using the*** `a_ex05_a.epx` ***file. After completing data entry, enter the same data again into the*** `a_ex05_b.epx` ***file.***
- ***After you have completed the two files, proceed to validation as explained here.***
- ***After ensuring that no record is missing in either file, export the*** `a_ex05_a.epx` ***file to a*** `a_ex05_f.epx` ***file, check out the discordances if any and correct them. This is your final dataset.***

On the next page you find the dataset with 15 records

Tuberculosis Programme Zimbabwe Form 2

Laboratory: Awuna

Tuberculosis laboratory register

Year: 2003

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