Exercise 4: 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 REC 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.

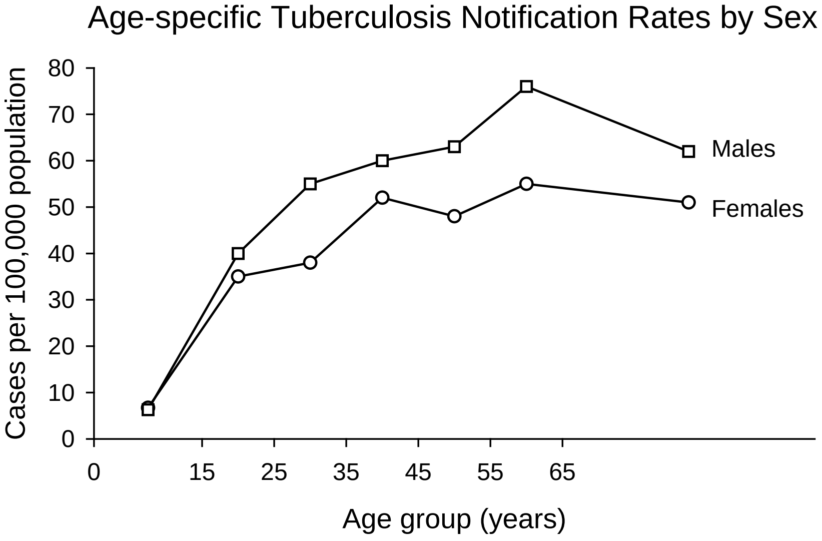
Ensuring quality data entry

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

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:

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

Using a *.CHK file

We have already a few inbuilt conditions that limit data entry errors by creating the A_EX03.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 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 against 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 data entry errors is miniscule.

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 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 (a *.not file) should be written. This can be invoked with **F5**. 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.

At this point in time, you will be using the first approach (we recommend this approach given its strength of enabling perfect documentation for future audits), and that is to independently enter the 15 records twice and then to compare the two files.

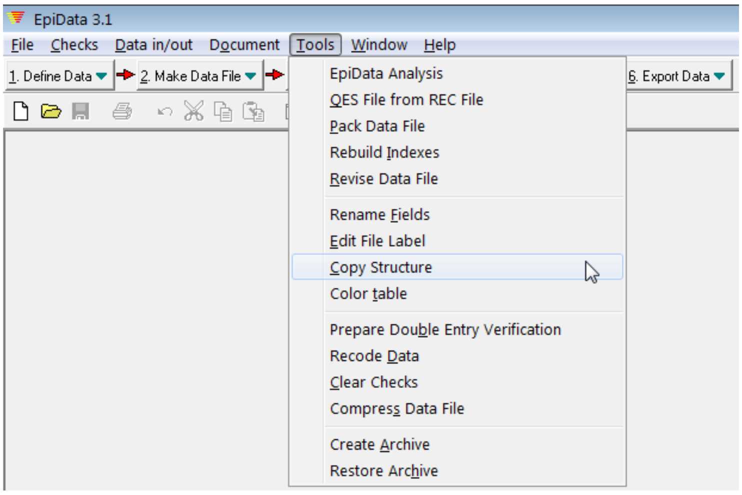
> Note for data entry: Do never move around the fields with the help of the mouse. The mouse movements cannot be recorded properly and unforeseen errors may occur (e.g., bypassing a calculation made in a field, missing a MUSTENTER command, etc), because the Check file cannot be applied to fields you 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.

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

Make a duplicate of the REC and CHK file pair

As we are entering the same data twice, we need two pairs of REC / CHK files, one for the first, the other for the second entry.

In “Tools” you find “Copy structure”:

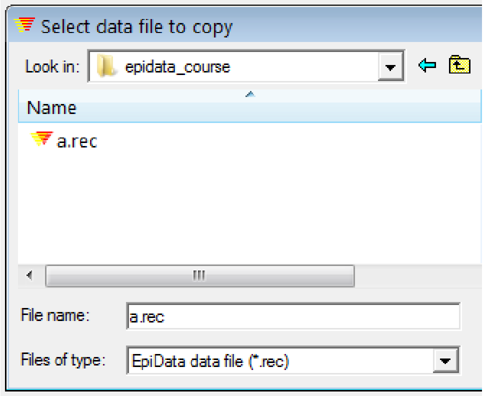


This allows making a copy of a REC file together with its CHK file to a new empty pair of files without copying the data. Thus, if you have A.REC and A.CHK and you use this feature, you can copy both the REC file and the CHK in one sweep to an empty B.REC and B.CHK file pair.

Tasks:

- ***Download the solution of Exercise 3 and overwrite your* A_EX03 *triplet files. Go to* "Tools" "Copy structure" *and copy the* A_EX03.REC *including its* A_EX03.CHK *file to:***

  A_EX04 A.REC ***and*** A_EX04A.CHK ***files***



Select data file to copy

Look in: epidata_course

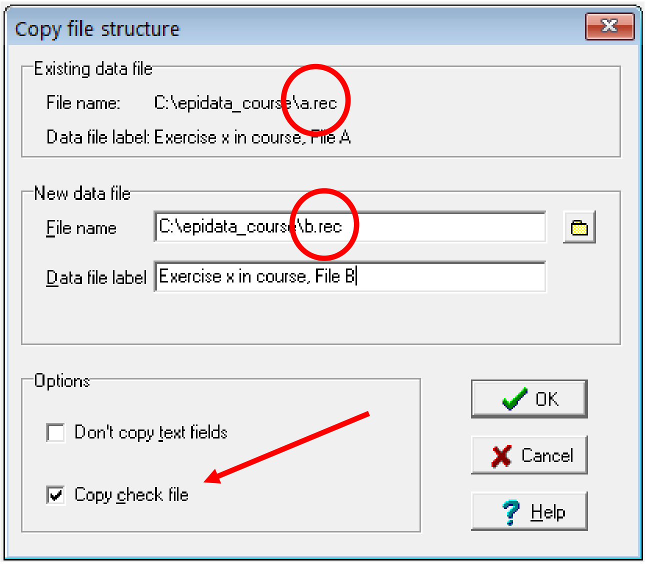
Name

a.rec

File name: a.rec

Files of type: EpiData data file (*.rec)

After selecting the A.REC file you get to the menu that allows you to give the desired name (here B.REC) to the New data file and keep the default Copy check file ticked:



Copy file structure

Existing data file

File name: C:\epidata_course\a.rec

Data file label: Exercise x in course, File A

New data file

File name C:\epidata_course\b.rec

Data file label Exercise x in course, File B

Options

☐ Don't copy text fields

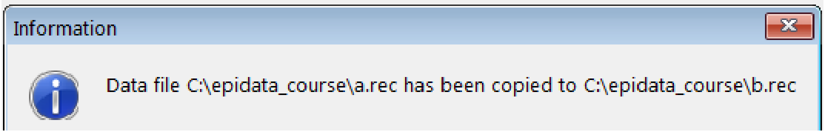
☑ Copy check file

OK

Cancel

Help

Both new files are created at the same time and you get confirmation when you enter:



Information

Data file C:\epidata_course\a.rec has been copied to C:\epidata_course\b.rec

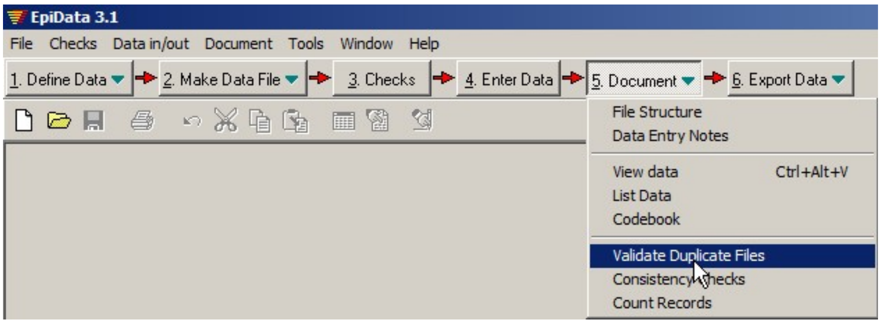
Again, note that this is for copying the structure (empty REC file and associated CHK file), not exporting data.

Double-entry

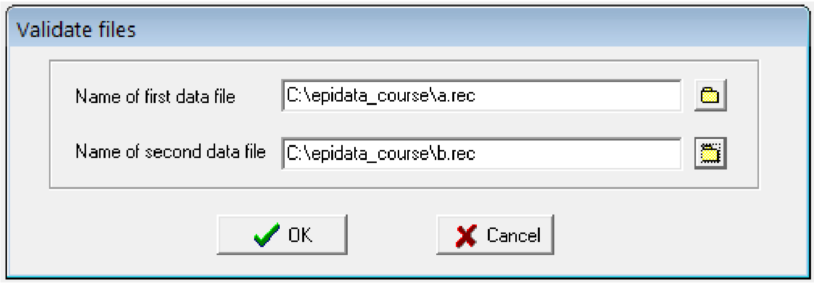
After you have both file pairs, enter the data into the A.REC (controlled by the A.CHK) and when completed, repeat data entry into the B.REC (controlled by the B.CHK).

Data validation

After completing double-entry, the two data files are compared, a procedure termed "**data validation**". In the process bar, go to 5. Document and choose Validate Duplicate Files:



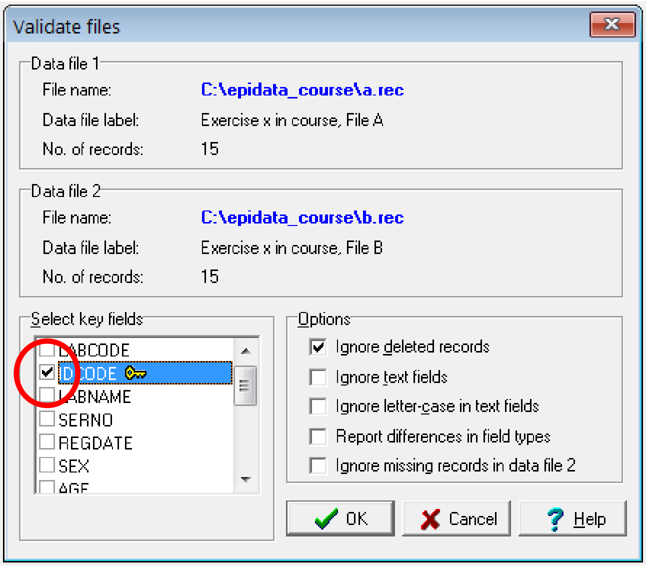
Enter the names of the two files that you need to compare into the respective two lines:



Name of first data file: C:\epidata_course\a.rec

Name of second data file: C:\epidata_course\b.rec

Confirming with OK opens:



Data file 1
File name: C:\epidata_course\a.rec
Data file label: Exercise x in course, File A
No. of records: 15

Data file 2
File name: C:\epidata_course\b.rec
Data file label: Exercise x in course, File B
No. of records: 15

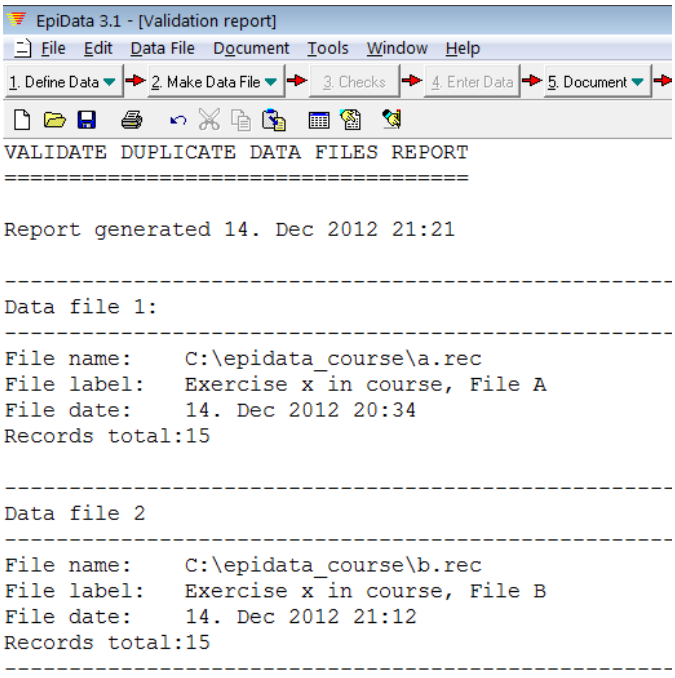
Select key fields
□ LABCODE
☑ IDCODE
□ LABNAME
□ SERNO
□ REGDATE
□ SEX
□ AGE

Options
☑ Ignore deleted records
□ Ignore text fields
□ Ignore letter-case in text fields
□ Report differences in field types
□ Ignore missing records in data file 2

We must compare the two files on “something”, and that something is our unique identifier `IDCODE` which we assured to be unique with `KEY UNIQUE`, and has therefore a little key to the right of the field name. We tick this field. This means that the `IDCODE` of a given record in File A determines that EpiData looks at the record with the same `IDCODE` in File B, irrespective of where in the sequence of records in the dataset it is found. This makes the sequence of entry of records irrelevant.

The Validation report

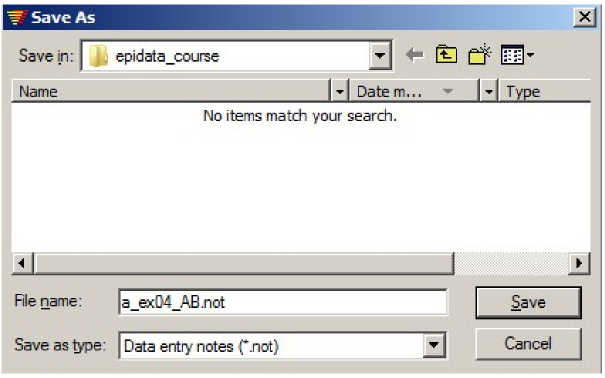
Once you OK, EpiData produces the Validation report:


EpiData 3.1 - [Validation report]
File Edit Data File Document Tools Window Help
1. Define Data
2. Make Data File
3. Checks
4. Enter Data
5. Document

```
VALIDATE DUPLICATE DATA FILES REPORT
====================================

Report generated 14. Dec 2012 21:21

-----------------------------------------------------
Data file 1:
-----------------------------------------------------
File name:    C:\epidata_course\a.rec
File label:   Exercise x in course, File A
File date:    14. Dec 2012 20:34
Records total:15

-----------------------------------------------------
Data file 2
-----------------------------------------------------
File name:    C:\epidata_course\b.rec
File label:   Exercise x in course, File B
File date:    14. Dec 2012 21:12
Records total:15
-----------------------------------------------------
```

Verify first that 1) the number of records is the same in both files and 2) that nowhere a record is present in File A but missing in File B and vice versa. Should you find missing records, you must add them first before repeating validation. Once you have assured that both datasets have the same records, you must save the Validation report. This is not only essential for subsequent corrections (if needed), but to have a permanent record that you did indeed validate the two files and what the results were. This is of critical importance.

Validation files are simple text files with the extension `*.NOT`. Give it a name that makes it intuitively clear what this validation refers to, like `*_AB.NOT`:



Save in: epidata_course

No items match your search.

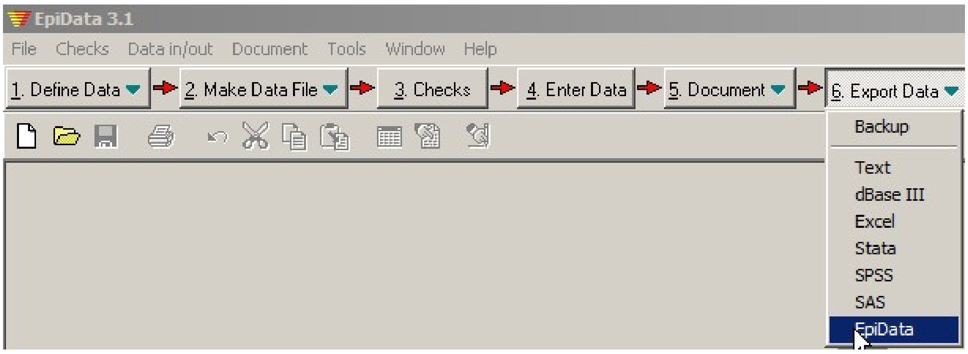
File name: a_ex04_AB.not

Save as type: Data entry notes (*.not)

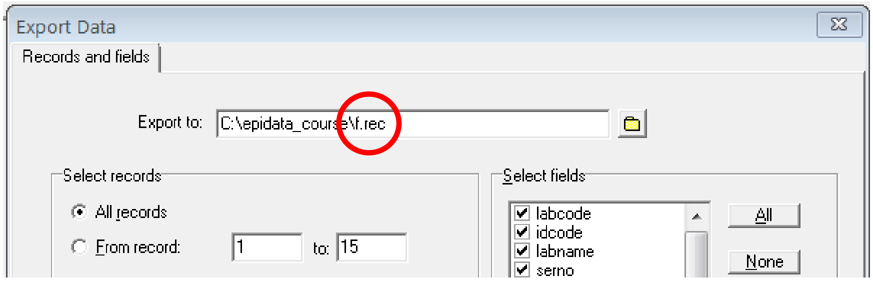
Creating a final dataset

One might be tempted to make corrections of any errors that might be identified through discordances in either the A.REC or the B.REC file. Doing so would, however, violate 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 F.REC file. To standardize as many things as possible, we always export the A.REC file to the F.REC file (even if in fact it is irrelevant whether we use the A.REC or the B.REC file, but consistency is good policy).

Go to “6. Export Data” | “Epidata”:



and export the A.REC file to the F.REC file:



Export to: C:\epidata_course\f.rec

Select records

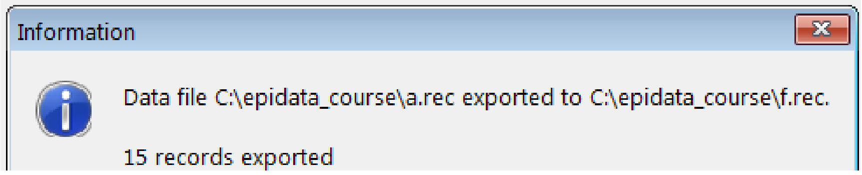
● All records

○ From record: 1 to: 15

Select fields

☑ labcode
☑ idcode
☑ labname
☑ serno

You get confirmation:



Data file C:\epidata_course\a.rec exported to C:\epidata_course\f.rec.

15 records exported

Note: While this process of exporting data is perfectly fine for exporting the dataset that we created, we find that there are some errors (detailed in the next chapter) in the export of the CHK file. Unfortunately, this is due to **a bug in EpiData** during the export of the CHK file. The EpiData Association is conscious of these issues (which have been discovered only after work on the new EpiData Manager started in earnest). There are no plans to fix them because of the current work on the revision of EpiData software which are well on the way. The

EpiData Manager will combine the EpiData triplet of QES, REC and CHK files and have a different architecture altogether. Every effort will be made to ensure bug-free functionality in data export from the current version to the new version.

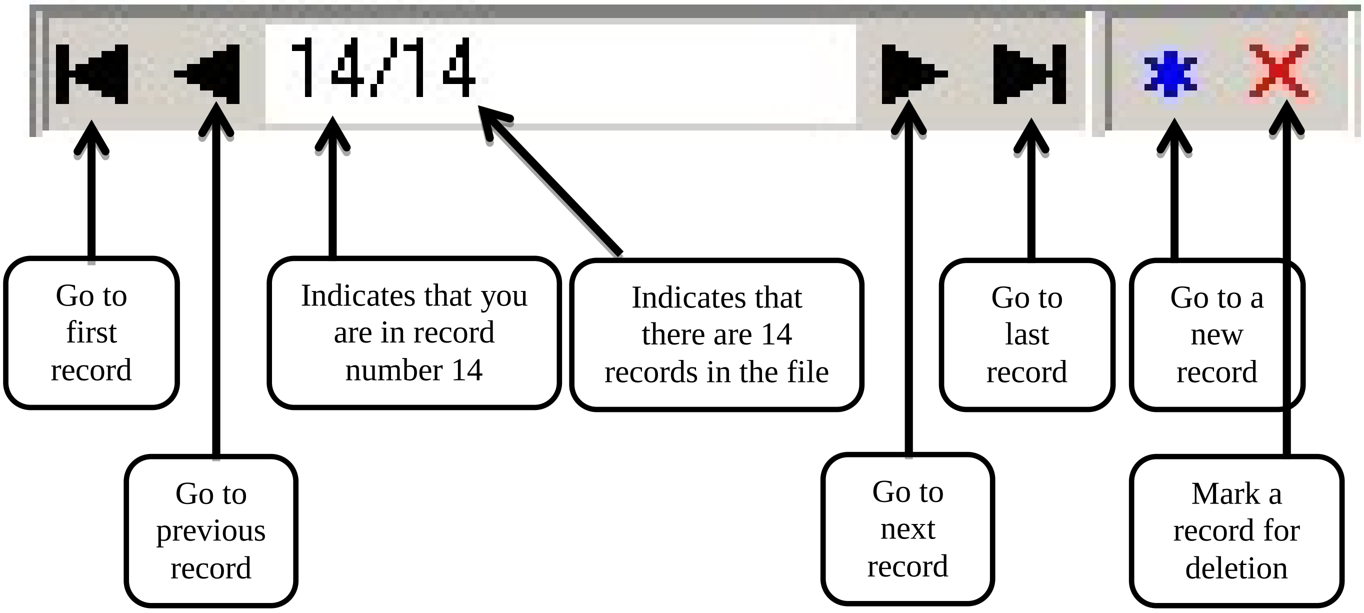
Hence, we recommend an alternative option as a workaround for the time being. Just create a copy of the REC and CHK files and rename them appropriately!

How to navigate through a REC file?

Before you begin, a few more tips on navigating through the REC file and manipulating them will help you in this exercise.

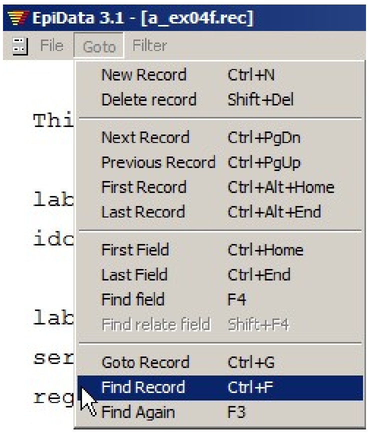
To navigate through different fields of a record, we recommend to use the ENTER key or ARROW keys or TAB key. Avoid using mouse since there is a possibility that checks are not applied when you use the mouse.

To navigate between the records of the REC file use the navigation panel on the left bottom end of the data entry screen which can be used to navigate through the records (see the diagram below to understand the different icons):

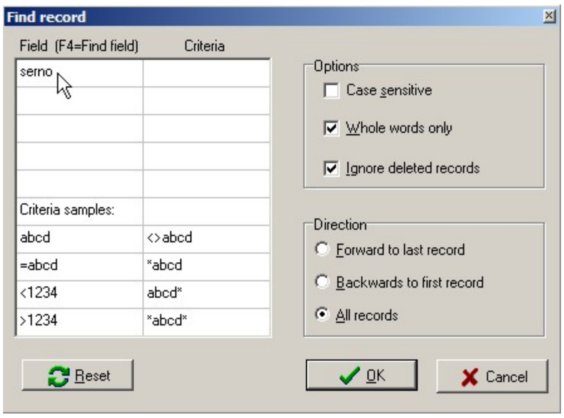

14/14
Go to first record
Go to previous record
Indicates that you are in record number 14
Indicates that there are 14 records in the file
Go to next record
Go to last record
Go to a new record
Mark a record for deletion

If you need to find a specific record and know the record number, use the function GOTO record **(CTRL+G).** If you do not know the record number, but want to find a record by specifying some criteria, use the function FIND record **(CTRL+F).** Let us demonstrate the latter option.

Let us say you want to find a record with the record with the unique identifier ML_J-2003-3303. Then click on GOTO in the menu and choose “Find Record”.



Click on it to find the dialogue box below



Find record

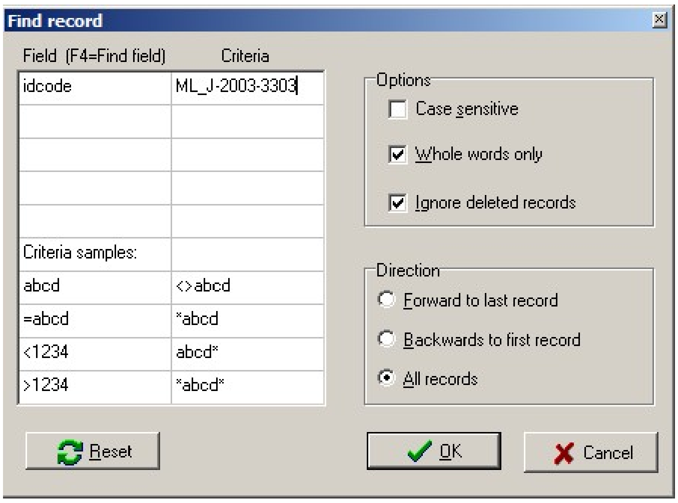
Field (F4=Find field)	Criteria
serno	
Criteria samples:	
abcd	<>abcd
=abcd	*abcd
<1234	abcd*
>1234	*abcd*

Options
- ☐ Case sensitive
- ☑ Whole words only
- ☑ Ignore deleted records

Direction
- ○ Forward to last record
- ○ Backwards to first record
- ● All records

Note that you can get to the same place using the shortcut (**CTRL+F**).

Now, edit the Field and the Criteria as shown below:



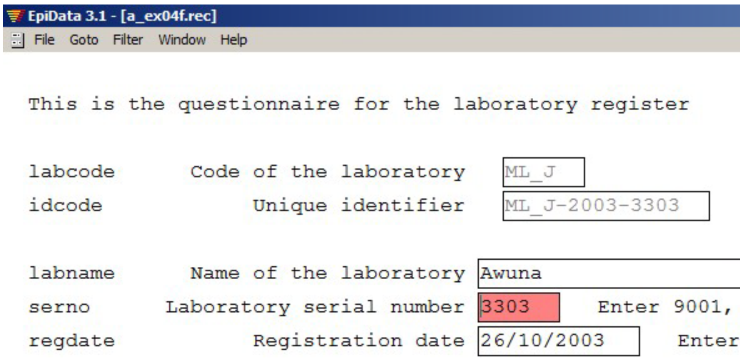
Find record

Field (F4=Find field)	Criteria
idcode	ML_J-2003-3303
Criteria samples:	
abcd	<>abcd
=abcd	*abcd
<1234	abcd*
>1234	*abcd*

Options
- ☐ Case sensitive
- ☑ Whole words only
- ☑ Ignore deleted records

Direction
- ○ Forward to last record
- ○ Backwards to first record
- ● All records

Click OK and then you will find yourself in the record with the sought after identifier:



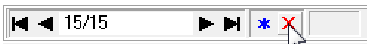
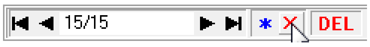
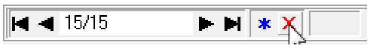
```
This is the questionnaire for the laboratory register

labcode      Code of the laboratory  ML_J
idcode            Unique identifier  ML_J-2003-3303

labname      Name of the laboratory Awuna
serno     Laboratory serial number 3303      Enter 9001,
regdate           Registration date 26/10/2003     Enter
```

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.

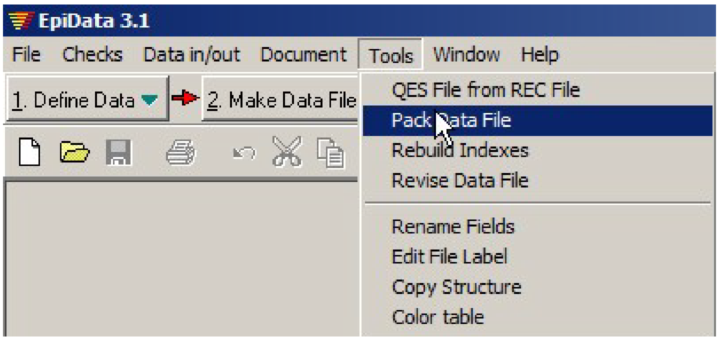
Steps in marking a record for deletion (Look at the screenshot below)

1. Open the REC 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:

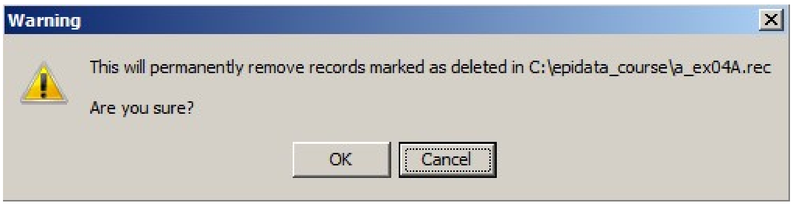
Click on the red cross	DEL appears	Click on the red cross again and DEL disappears
		

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

First go to Tools, select Pack Data File and click on it.



You will receive a warning about permanent deletion of the record:

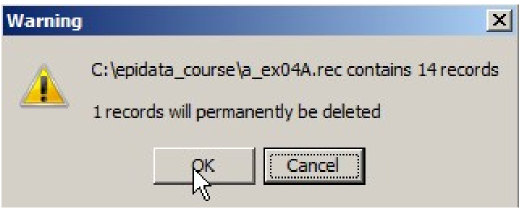


Warning

This will permanently remove records marked as deleted in C:\epidata_course\a_ex04A.rec

Are you sure?

If you press OK you will again receive another warning – this time more specifically mentioning how many records will be deleted. We hope you appreciate the effort of EpiData to protect your data!

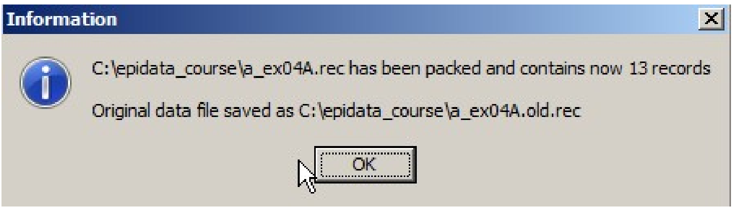


Warning

C:\epidata_course\a_ex04A.rec contains 14 records

1 records will permanently be deleted

If you click OK now, this will delete the records marked for deletion. But EpiData will preserve the original REC file as a_ex04A.old.rec in case you want to revert back!



Information

C:\epidata_course\a_ex04A.rec has been packed and contains now 13 records

Original data file saved as C:\epidata_course\a_ex04A.old.rec

You should now be prepared for the task of entering, double-entering, and validating the data

Tasks:

- ***Download the solution of Exercise 3 and overwrite your*** A_EX03 ***triplet files. Go to*** "Tools" "Copy structure" ***and copy the*** A_EX03.REC ***including its*** A_EX03.CHK ***file to:***

  A_EX04 A.REC ***and*** A_EX04A.CHK ***files***

  A_EX04 B.REC ***and*** A_EX04B.CHK ***files***

- ***Enter the 15 records using the*** A_EX04A.REC ***file. After completing data entry, enter the same data again into to the*** A_EX04B.REC ***file.***

- ***After you have completed the two files, go to*** "5. Document" "Validate Duplicate Files" ***and produce a*** *.NOT ***file giving you a list of discordances, if any. Save the*** *.NOT ***file as*** A_EX04AB_validation.NOT

- ***Use*** "6. Export Data" "Epidata" ***to export either one of the two files [recommended: the*** A_EX04A.REC***] to a new*** A_EX04F.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

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 or name of the referral BMU (Basic Management Unit)