

Part C. Operations Research

Part C: Operations research

Exercise 1: Creating a working dataset

Exercise 2: Variability in serial smears

Exercise 3: Incremental yield from serial smears

Exercise 4: Confirmatory results in serial smears

Introductory note

In this Part C three operationally relevant research questions will be answered:

- Does the dataset tell us something about how diligent the work was performed in the tuberculosis microscopy laboratory?
- Is the third serial smear examination associated with an excessive amount of work for little gain?
- Is it necessary to confirm a positive smear result?

These and related questions were asked by graduates from The Union's operations research courses in fulfillment of the field component of the course. The data were collected in Moldova (Dr Dumitru Laticevschi, fifth course, Paris, 2003), Mongolia (Dr Nymadawaa Naranbat, seventh course, Paris, 2004), Uganda (Dr Achilles Katamba, fifth course, Paris, 2003), and Zimbabwe (Dr Biggie Mabaera, seventh course, Paris, 2004). Six publications have resulted from this study:

Mabaera B, Naranbat N, Dhliwayo P, Rieder H L. Efficiency of serial smear examinations in excluding sputum smear-positive tuberculosis. *Int J Tuberc Lung Dis* 2006;10:1030-5.

Katamba A, Laticevschi D, Rieder H L. Efficiency of a third serial sputum smear examination in the diagnosis of tuberculosis in Moldova and Uganda. *Int J Tuberc Lung Dis* 2007;11:659-64.

Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Sputum smear-positive tuberculosis: empiric evidence challenges the need for confirmatory smears. *Int J Tuberc Lung Dis* 2007;11:959-64.

Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Making pragmatic sense of data in the tuberculosis laboratory register. *Int J Tuberc Lung Dis* 2008;12:294-300.

Mabaera B, Naranbat N, Katamba A, Laticevschi D, Lauritsen J M, Rieder H L. Seasonal variation among tuberculosis suspects in four countries. *International Health* 2009;1:53-60.

Rieder H L, Lauritsen J M, Naranbat N, Katamba A, Laticevschi D, Mabaera B. Quantitative differences in sputum smear microscopy results for acid-fast bacilli by age and sex in four countries. *Int J Tuberc Lung Dis* 2009;13:1393-8.

With permission of the investigators, the datasets have been made publicly accessible for use in this course exactly as they have been collected.

Exercise 1: Creating a working dataset

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

- a. Combine different datasets into one combined dataset
- b. Recode 'text variables' to 'numeric variables'
- c. Remove 'undesirable' records from a dataset
- d. Correct obvious gross errors from the datasets
- e. Create a 'cleaned' final working dataset from available datasets

Moldova and Uganda worked together using the same data entry forms. You obtained MOL_25.ZIP and UGA_30.ZIP. These two files contain respectively the data files obtained from the 25 laboratories in Moldova and the data files obtained from the 30 laboratories in Uganda. In addition, each of the zip files contains the base pair of QES and CHK files (which are identical for both countries).

Mongolia and Zimbabwe worked together using the same data entry forms. You obtained MON_31.ZIP and ZIM_23.ZIP. These two files contain respectively the data files obtained from the 31 laboratories in Mongolia and the data files obtained from the 23 laboratories in Zimbabwe. In addition, each of the zip files contains the base pair of QES and CHK files (which are identical for both countries).

The two pairs of countries collected exactly the same information from the laboratory register, but their data collection forms (the QES files, and thus REC files) and CHK files had small differences. You can find these by inspecting the respective files. However, as you come in here as an outsider, we summarize these in the following table, and also give you the field names that the final data set combining all files should have.

| Field label | Field name Moldova / Uganda | Field name Mongolia / Zimbabwe | Final Field name |
|------------------------------|-----------------------------|--------------------------------|------------------|
| Study country | -- | -- | country |
| Laboratory code | labcode | laboratory | laboratory |
| Laboratory serial number | serno | serno | -- |
| Registration date | labdate | regdate | regdate |
| Year of registration | -- | -- | regyear |
| Created unique identifier | unique | Id | -- |
| Sex of examinee | sex | sex | sex |
| Age (in years) of examinee | age | age | age |
| Reason for examination | reason | reason | reason |
| Result of first examination | res1 | res1 | result1 |
| Result of second examination | res2 | res2 | result2 |
| Result of third examination | res3 | res3 | result3 |

Omissions and commissions

In contrast to what you learned in Part A, the data entry form used only field names but had no field labels.

In both studies SEX and REASON were coded as text fields rather than numerically and using label blocks. The fields RES1, RES2, and RES3 also differed slightly: a value of 4.0 did not exist in Moldova / Uganda, but denoted “Positive, not quantified” in Mongolia / Zimbabwe, while “Positive, not quantified” was coded as 8.0 in the latter but did not exist in the former. “Scanty, not quantified” was coded as 5.0 in Mongolia / Zimbabwe, but was forgotten as a possible value in Moldova / Uganda.

You could obtain the information from the CHK files, but the summary of the coding for the fields of relevance with the differences is as follows:

| Field name | Field value Moldova / Uganda | Field value Mongolia / Zimbabwe | Value label |
|------------------------|---|--|---|
| sex | F M 9 | F M 9 | Female Male Unknown sex |
| reason | D F 9 -- -- -- -- -- -- -- | D F 9 1 2 3 4 5 6 7 8 | Diagnosis Follow-up, month not stated Reason not stated Follow-up at 1 month Follow-up at 2 months Follow-up at 3 months Follow-up at 4 months Follow-up at 5 months Follow-up at 6 months Follow-up at 7 months Follow-up at 8 months or later |
| res1 (also res2, res3) | 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 2.0 3.0 -- -- 8.0 9.0 | 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 2.0 3.0 4.0 5.0 -- 9.0 | Negative Scanty, 1 AFB / 100 fields Scanty, 2 AFB / 100 fields Scanty, 3 AFB / 100 fields Scanty, 4 AFB / 100 fields Scanty, 5 AFB / 100 fields Scanty, 6 AFB / 100 fields Scanty, 7 AFB / 100 fields Scanty, 8 AFB / 100 fields Scanty, 9 AFB / 100 fields 1+ positive 2+ positive 3+ positive Positive, not quantified Scanty, not quantified Positive, not quantified No result recorded |

Tasks:

- o Create a combined dataset C_EX01_COMBINE.REC from all 107 files with a program C_EX01_COMBINE.PGM.*

Notes to the first task:

From the dataset from Moldova, drop the data for the laboratory “BND” (containing data from only 1 week) and remove one empty record.

From the dataset from Mongolia, remove the empty records

In Zimbabwe, one record has no laboratory value, but it has an ID (this is most likely attributable to some manipulation with the mouse after ID creation). You can retain this record by giving the laboratory the correct code that we know from the ID.

If you have removed all empty records (plus the one laboratory from Moldova) and you make a frequency of COUNTRY you should get:

| country | | |
|----------------|----------|----------|
| | N | % |
| MOL | 17865 | 13.7 |
| MON | 22588 | 17.3 |
| UGA | 55114 | 42.3 |
| ZIM | 34744 | 26.7 |
| Total | 130311 | 100.0 |

- o Create a “cleaned” final working dataset C_EX01.REC with a program C_EX01.PGM which excludes non-sensically coded result sequences, and with all fields codes numerically (including COUNTRY and LABORATORY).*

Notes to the second task:

For the numeric coding of the COUNTRY follow the alphabet: 1 for Moldova, 2 for Mongolia, ..., 4 for Zimbabwe.

For the numeric coding of the laboratories, make a frequency for each country, and then code numerically following the country notation:

Moldova laboratories:

```
if laboratory="ANR" then lab0=101
if laboratory="BLM" then lab0=102
if laboratory="BRL" then lab0=103
if laboratory="BSR" then lab0=104
if laboratory="CCE" then lab0=105
...etc
```

Mongolia laboratories:

```
if laboratory="AR_B" then lab0=201
if laboratory="BG_B" then lab0=202
if laboratory="BN_B" then lab0=203
...etc
```

Uganda laboratories:

```
if trim(laboratory)="1" then lab0=301
if trim(laboratory)="2" then lab0=302
if trim(laboratory)="3" then lab0=303
...etc
```

Zimbabwe laboratories:

```
if laboratory="BY_A" then lab0=401
if laboratory="MC_A" then lab0=402
if laboratory="MC_B" then lab0=403
if laboratory="MC_C" then lab0=404
...etc
```

We also propose to correct some obvious gross errors (which are obvious from the sequence in recording what they should have been) in the registration date. In order to get a common ground, we point these out here and provide the program file commands for these (note that we made a date variable just for this manipulation here):

```
define regyear0 ####
regyear0=year(regdate)
define regyear ####
regyear=regyear0
* correct errors in year of recording
if regyear0=1990 and laboratory=301 then regyear=1999
if regyear0=1990 and laboratory=306 then regyear=1999
if regyear0=1990 and laboratory=319 then regyear=1999
if regyear0=1990 and laboratory=320 then regyear=2000
if regyear0=1990 and laboratory=410 then regyear=2002

if regyear0=2000 and laboratory=408 then regyear=2002
if regyear0=2000 and laboratory=416 then regyear=2002
if regyear0=2000 and laboratory=419 then regyear=2002

if regyear0=2004 and laboratory=211 then regyear=2003
if regyear0=2004 and laboratory=223 then regyear=2003
if regyear0=2004 and laboratory=401 then regyear=2002
if regyear0=2004 and laboratory=408 then regyear=2002
if regyear0=2004 and laboratory=412 then regyear=2003
if regyear0=2004 and laboratory=413 then regyear=2002

if regyear0=2005 and laboratory=207 then regyear=2003
if regyear0=2033 and laboratory=207 then regyear=2003
```

If you have cleaned the dataset and you make a table of COUNTRY by REGYEAR you should get:

| Study country | | | | | |
|----------------------|---------|----------|--------|----------|--------|
| Year of registration | Moldova | Mongolia | Uganda | Zimbabwe | Total |
| 1999 | 0 | 0 | 17308 | 0 | 17308 |
| 2000 | 0 | 0 | 18655 | 0 | 18655 |
| 2001 | 0 | 0 | 18087 | 1213 | 19300 |
| 2002 | 0 | 149 | 0 | 29307 | 29456 |
| 2003 | 17725 | 22406 | 0 | 3958 | 44089 |
| Total | 17725 | 22555 | 54050 | 34478 | 128808 |

Note the following on the CHK and QES files:

If you start with a REC file that is accompanied by its CHK file and then create new variables with Field values and Value labels using the LABELVALUE, EpiData Analysis takes the original CHK file and appends it with the new Field values and their Value labels when you create a new REC file. You can also define a Field label (command LABEL newvar "X"). However, Field labels are not part of the CHK file, they come from the QES file, integrated into the REC file.

You can back-create a QES file from a REC file, but it will be missing some formatting, such as aligned fields or sequence of fields. As you know from Part A, whenever you open a REC file in EpiData Entry, it checks whether there is a newer QES file, and if found, ask whether you want to adapt the REC file to that QES file.

Create this way a QES file (C_EX01.QES) from the C_EX01.REC file, then open it and you may note that it made automatic field naming:

```
Age                ##
{country} Study country      #
{regdate} Date of registration <dd/mm/yyyy>
{Result} of {1}st examination #.#
{Result} of {2}nd examination #.#
{Result} of {3}rd examination #.#
{Reason} for examination    ##
{Sex} of examinee          #
{Laboratory} code          ###
{regyear} Year of registration #####
```

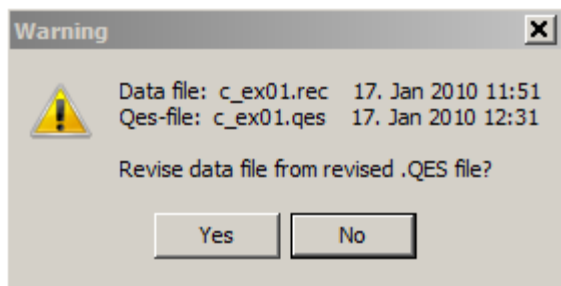
Edit it properly (and change to first-word field naming) to get:

```
Laboratory register study
Mongolia, Moldova, Uganda, Zimbabwe
```

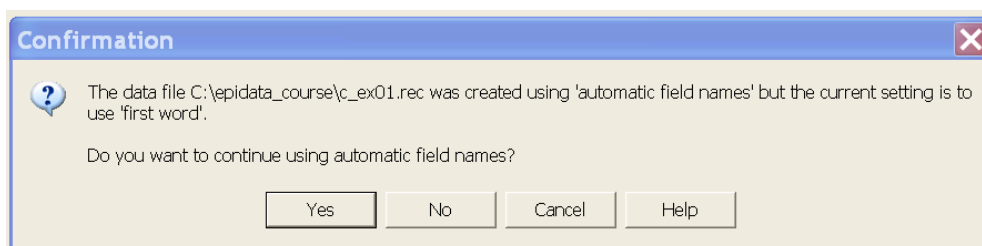
Date is displayed in day-month-year format

```
country           Study country #
laboratory        laboraotry code ###
regdate           Date of registration <dd/mm/yyyy>
regyear           Year of registration #####
age               Patient's age in years ##
sex               Sex of examinee #
reason            Reason for examination ##
result1           Result of 1st examination #.#
result2           Result of 2nd examination #.#
result3           Result of 3rd examination #.#
```

Open the C_EX01.QES file and make a non-changing change (e.g., a hard carriage return, followed by undoing it) and save the C_EX01.QES (which is the same as the original). Then try to enter data and you get the following message:



After you confirm, you get:



As we are using First word, not automatic field names, you must choose "No". After that the REC file opens, and if you go to the last completed record, you get:

Laboratory Register Study
Mongolia, Moldova, Uganda, Zimbabwe

Date is displayed in Day-Month-Year format

| | | |
|------------|---------------------------|------------|
| country | Study country | 4 |
| laboratory | Laboratory code | 402 |
| regdate | Registration date | 16/03/2003 |
| regyear | Year of registration | 2003 |
| age | Age in years | 99 |
| sex | Sex of examinee | 2 |
| reason | Reason for examination | 0 |
| result1 | Result of 1st examination | 0.0 |
| result2 | Result of 2nd examination | 1.0 |
| result3 | Result of 3rd examination | 9.0 |