

## Solution to Exercise 1: A relational database and “Aggregating” vs from “long-to-wide”

### Key points:

- A relational database is the solution to a varying number of observations per individual
- The child file is merged with the parent file to give a dataset of all observations
- To obtain means for an individual from continuous variables, aggregating the data is the strategy of choice
- To reduce the dataset to individuals with information on each examination, one must copy the information from observations in the vertical to newly created fields in the first observation of each individual before selecting that record from the individual (“long-to-wide”)

### Task

- *Prepare a data documentation sheet*

The documentation sheet is shown on the next page.

### Task

- *Prepare two EpiData Entry triplets to create a relational database*

The data entry forms may be made as follows:

D\_EX01\_PATIENT.QES

Data entry form: Patient information

patid	Unique patient identifier	<input type="text"/>
sex	Patient's sex	<input type="text"/>
marital	Patient's marital status	<input type="text"/>

D\_EX01\_EXAMINATION.QES

Data entry form: Examination information

patid	Unique patient identifier	<input type="text"/>
examid	Unique examination identifier	<input type="text"/>
dateexam	Date of examination	<input type="text"/> Enter 01/01/1800 if missing
bs	Fasting plasma blood glucose (mMol)	<input type="text"/> Enter 99.9 if missing
sputum	Macroscopic sputum aspect	<input type="text"/>
micro	Microscopy result	<input type="text"/>

The data documentation sheet:

### Data documentation sheet

	Field name	Field label	Field type	Field length	Field values	Value label	Field comment
<b>Patient file</b>	patid	Unique patient identifier	U	1	A,....,Z		Any given unique ID
	sex	Patient's sex	I	1		1 Female 2 Male 3 Unknown	
	marital	Patient's marital status	I	1		1 Annulled 2 Cohabiting 3 Divorced 4 Engaged 5 Married 6 Separated 7 Single 8 Widowed 9 Unknown	

NOTE: If SEX is given during an earlier examination and left empty in a subsequent examination, keep information from earlier  
 If SEX is different in different examinations, record as UNKNOWN  
 If MARITAL is given during an earlier examination and left empty in a subsequent examination, keep information from earlier  
 If MARITAL is different in different examinations, update to most recent information

<b>Examination file</b>	examid	Unique examination identifier	S	12	A-2007-01-31,...		Automatically calculated
	datexam	Date of examination	dd/mm/yyyy	10	01/01/2007,....,31/12/2007		Legal visit date recordings
	bs	Fasting plasma blood glucose (mMol)	F	4	2.5,....,19.9		Legal valid value Enter if blood sugar is missing

sputum	Macroscopic sputum aspect	I	1	<ul style="list-style-type: none"> <li>1 Mucoid</li> <li>2 Purulent</li> <li>3 Muco-purulent</li> <li>4 Blood-tinged</li> <li>5 Salivary</li> <li>6 Other</li> <li>9 Unknown</li> </ul>
result	Examination result	F	3	<ul style="list-style-type: none"> <li>0 Negative</li> <li>1 "1+ positive"</li> <li>2 "2+ positive"</li> <li>3 "3+ positive"</li> <li>9 "No result recorded"</li> <li>4 "Positive, not quantified"</li> <li>5 "Scanty, not quantified"</li> <li>0.1 "Scanty, 1 AFB per 100 fields"</li> <li>0.2 "Scanty, 2 AFB per 100 fields"</li> <li>0.3 "Scanty, 3 AFB per 100 fields"</li> <li>0.4 "Scanty, 4 AFB per 100 fields"</li> <li>0.5 "Scanty, 5 AFB per 100 fields"</li> <li>0.6 "Scanty, 6 AFB per 100 fields"</li> <li>0.7 "Scanty, 7 AFB per 100 fields"</li> <li>0.8 "Scanty, 8 AFB per 100 fields"</li> <li>0.9 "Scanty, 9 AFB per 100 fields"</li> </ul>

The D\_EX01\_PATIENT.CHK file:

```
LABELBLOCK
  LABEL label_sex
    1 Female
    2 Male
    9 Unknown
  END
  LABEL label_marital
    1 Annulled
    2 Cohabiting
    3 Divorced
    4 Engaged
    5 Married
    6 Separated
    7 Single
    8 Widowed
    9 Unknown
  END
END

patid
  KEY UNIQUE 1
  MUSTENTER
END

sex
  COMMENT LEGAL USE label_sex SHOW
  MUSTENTER
  TYPE COMMENT
  AFTER ENTRY
  RELATE patid d_ex01_examination.rec
  END
END

marital
  COMMENT LEGAL USE label_marital SHOW
  MUSTENTER
  TYPE COMMENT
  END
```

Note:

The line:

```
RELATE patid d_ex01_examination.rec
```

is for the final file. It must be adapted accordingly for the two duplicate files which you make first for validation!

The D\_EX01\_EXAMINATION.CHK file

```
LABELBLOCK
  LABEL label_sputum
    1 Mucoid
    2 Purulent
    3 Muco-purulent
    4 Blood-tinged
    5 Salivary
```

```

        6 Other
        9 Unknown
    END
    LABEL label_result
        0.0 Negative
        1.0 "1+ positive"
        2.0 "2+ positive"
        3.0 "3+ positive"
        9.0 "No result recorded"
        4.0 "Positive, not quantified"
        5.0 "Scanty, not quantified"
        0.1 "Scanty, 1 AFB per 100 fields"
        0.2 "Scanty, 2 AFB per 100 fields"
        0.3 "Scanty, 3 AFB per 100 fields"
        0.4 "Scanty, 4 AFB per 100 fields"
        0.5 "Scanty, 5 AFB per 100 fields"
        0.6 "Scanty, 6 AFB per 100 fields"
        0.7 "Scanty, 7 AFB per 100 fields"
        0.8 "Scanty, 8 AFB per 100 fields"
        0.9 "Scanty, 9 AFB per 100 fields"
    END
END

patid
    KEY 1
    NOENTER
END

examid
    KEY 2
    NOENTER
END

dateexam
    RANGE 01/01/2007 31/12/2007
    LEGAL
        01/01/1800
    END
    MUSTENTER
    AFTER ENTRY
        examid=patid+"-"+month(dateexam)+"-"+day(dateexam)
    END
END

bs
    MUSTENTER
END

sputum
    COMMENT LEGAL USE label_sputum SHOW
    MUSTENTER
    TYPE COMMENT
END

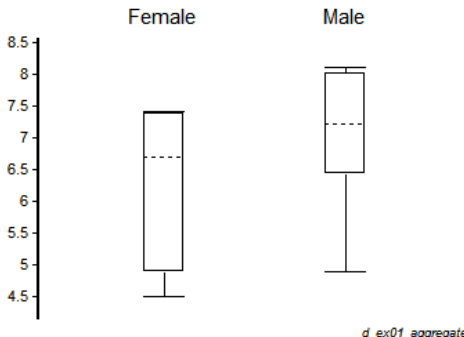
micro
    COMMENT LEGAL USE label_result SHOW
    MUSTENTER
    TYPE COMMENT
END

```

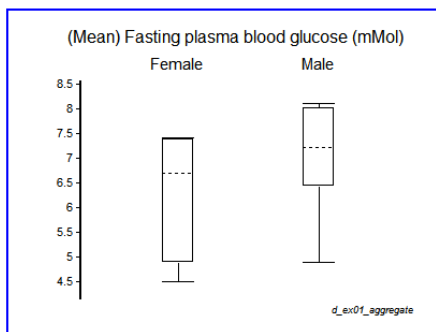
This file is the same for all three files.

**Tasks:**

- Write a program *D\_EX01.PGM* that merges the two files, then prepare sets for the aggregated data and for the “long-to-wide” transformation to produce the following output (obtaining the same numbers is relevant) respectively:

<p style="text-align: center;">From aggregating the data:</p> <p style="text-align: center;">(Mean) Fasting plasma blood glucose (mMol)</p>  <p style="text-align: right; font-size: small;">d_ex01_aggregate</p>	<p style="text-align: center;">From transformation “long-to-wide”:</p> <table border="1" style="width: 100%; border-collapse: collapse; font-size: small;"> <thead> <tr> <th colspan="4" style="text-align: center;">Overall microscopy result</th> </tr> <tr> <th style="text-align: left;">of 4 serial smears</th> <th style="text-align: center;">Negative</th> <th style="text-align: center;">Positive</th> <th style="text-align: center;">Total</th> </tr> </thead> <tbody> <tr><td>N—</td><td style="text-align: center;">2</td><td style="text-align: center;">0</td><td style="text-align: center;">2</td></tr> <tr><td>NN—</td><td style="text-align: center;">1</td><td style="text-align: center;">0</td><td style="text-align: center;">1</td></tr> <tr><td>NN9P</td><td style="text-align: center;">0</td><td style="text-align: center;">1</td><td style="text-align: center;">1</td></tr> <tr><td>NP—</td><td style="text-align: center;">0</td><td style="text-align: center;">1</td><td style="text-align: center;">1</td></tr> <tr><td>NPP—</td><td style="text-align: center;">0</td><td style="text-align: center;">3</td><td style="text-align: center;">3</td></tr> <tr><td>PNP—</td><td style="text-align: center;">0</td><td style="text-align: center;">1</td><td style="text-align: center;">1</td></tr> <tr><td>PP—</td><td style="text-align: center;">0</td><td style="text-align: center;">1</td><td style="text-align: center;">1</td></tr> <tr><td><b>Total</b></td><td style="text-align: center;"><b>3</b></td><td style="text-align: center;"><b>7</b></td><td style="text-align: center;"><b>10</b></td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse; font-size: small;"> <thead> <tr> <th colspan="5" style="text-align: center;">Patient's sex</th> </tr> <tr> <th style="text-align: left;">Incremental yield of first 3 smears</th> <th style="text-align: center;">Female</th> <th style="text-align: center;">% Male</th> <th style="text-align: center;">% Total</th> <th style="text-align: center;">%</th> </tr> </thead> <tbody> <tr><td>NNP</td><td style="text-align: center;">0 (0.0)</td><td style="text-align: center;">1 (25.0)</td><td style="text-align: center;">1 (14.3)</td><td></td></tr> <tr><td>NPx</td><td style="text-align: center;">3 (100.0)</td><td style="text-align: center;">1 (25.0)</td><td style="text-align: center;">4 (57.1)</td><td></td></tr> <tr><td>Px</td><td style="text-align: center;">0 (0.0)</td><td style="text-align: center;">2 (50.0)</td><td style="text-align: center;">2 (28.6)</td><td></td></tr> <tr><td><b>Total</b></td><td style="text-align: center;"><b>3 (100.0)</b></td><td style="text-align: center;"><b>4 (100.0)</b></td><td style="text-align: center;"><b>7</b></td><td></td></tr> </tbody> </table> <p style="font-size: x-small;">Percents: (Col)</p>	Overall microscopy result				of 4 serial smears	Negative	Positive	Total	N—	2	0	2	NN—	1	0	1	NN9P	0	1	1	NP—	0	1	1	NPP—	0	3	3	PNP—	0	1	1	PP—	0	1	1	<b>Total</b>	<b>3</b>	<b>7</b>	<b>10</b>	Patient's sex					Incremental yield of first 3 smears	Female	% Male	% Total	%	NNP	0 (0.0)	1 (25.0)	1 (14.3)		NPx	3 (100.0)	1 (25.0)	4 (57.1)		Px	0 (0.0)	2 (50.0)	2 (28.6)		<b>Total</b>	<b>3 (100.0)</b>	<b>4 (100.0)</b>	<b>7</b>	
Overall microscopy result																																																																							
of 4 serial smears	Negative	Positive	Total																																																																				
N—	2	0	2																																																																				
NN—	1	0	1																																																																				
NN9P	0	1	1																																																																				
NP—	0	1	1																																																																				
NPP—	0	3	3																																																																				
PNP—	0	1	1																																																																				
PP—	0	1	1																																																																				
<b>Total</b>	<b>3</b>	<b>7</b>	<b>10</b>																																																																				
Patient's sex																																																																							
Incremental yield of first 3 smears	Female	% Male	% Total	%																																																																			
NNP	0 (0.0)	1 (25.0)	1 (14.3)																																																																				
NPx	3 (100.0)	1 (25.0)	4 (57.1)																																																																				
Px	0 (0.0)	2 (50.0)	2 (28.6)																																																																				
<b>Total</b>	<b>3 (100.0)</b>	<b>4 (100.0)</b>	<b>7</b>																																																																				

To get:



"Table 1. Pattern of serial smear results"

definition by microscopy			
of 4 serial smears	Negative	Positive	Total
N—	2	0	2
NN—	1	0	1
NN9P	0	1	1
NP—	0	1	1
NPP—	0	3	3
PNP—	0	1	1
PP—	0	1	1
<b>Total</b>	<b>3</b>	<b>7</b>	<b>10</b>

"Table 2. Incremental yield among positive results"

Patient's sex				
Incremental yield of first 3 smears	Female	% Male	% Total	%
NNP	0 (0.0)	1 (25.0)	1 (14.3)	
NPx	3 (100.0)	1 (25.0)	4 (57.1)	
Px	0 (0.0)	2 (50.0)	2 (28.6)	
<b>Total</b>	<b>3 (100.0)</b>	<b>4 (100.0)</b>	<b>7</b>	

Percents: (Col)

The D\_EX01.PGM reads:

```
* Exercise D_EX01
* Merging files and aggregating files
* Copying and transposing data from "Long-to-wide"

cls
close
logclose

read "d_ex01_examination.rec"
merge patid /file="d_ex01_patient.rec" /table

sort patid dateexam
gen i exam=1
if patid=patid[_n-1] then exam=exam[_n-1]+1
label exam "Number of examination"

savedata "templ.rec" /replace

cls
close
read "templ.rec"

*****
* Create an aggregate data set
* to determine means

aggregate patid sex /mean="bs" /save="d_ex01_aggregate.rec" /replace

cls
close
read "d_ex01_aggregate.rec"

* set display databrowser=on
* browse
* tables sex meabs
* boxplot meabs /by=sex
*****
* Transpose / copy "long-to-wide"

cls
close
read "templ.rec"

* freq exam

cls
gen d dateexam1=date("31/12/1899")
gen d dateexam2=date("31/12/1899")
gen d dateexam3=date("31/12/1899")
gen d dateexam4=date("31/12/1899")
                                dateexam1=dateexam
if (patid[_n])=(patid[_n+1]) then dateexam2=dateexam[_n+1]
if (patid[_n])=(patid[_n+2]) then dateexam3=dateexam[_n+2]
if (patid[_n])=(patid[_n+3]) then dateexam4=dateexam[_n+3]

cls
gen f micro1=-1
gen f micro2=-1
gen f micro3=-1
gen f micro4=-1
                                micro1=micro
if (patid[_n])=(patid[_n+1]) then micro2=micro[_n+1]
```

```

if (patid[_n])=(patid[_n+2]) then micro3=micro[_n+2]
if (patid[_n])=(patid[_n+3]) then micro4=micro[_n+3]

cls
gen i sputum1=-1
gen i sputum2=-1
gen i sputum3=-1
gen i sputum4=-1

                                sputum1=sputum
if (patid[_n])=(patid[_n+1]) then sputum2=sputum[_n+1]
if (patid[_n])=(patid[_n+2]) then sputum3=sputum[_n+2]
if (patid[_n])=(patid[_n+3]) then sputum4=sputum[_n+3]

select exam=1
savedata "temp2.rec" /replace

cls
close
read "temp2.rec"

define res1 ##
res1=integer(micro1)*10
label res1 "Microscopy result of 1st examination"

cls
define res2 ##
res2=integer(micro2)*10
label res2 "Microscopy result of 2nd examination"

cls
define res3 ##
res3=integer(micro3)*10
label res3 "Microscopy result of 3rd examination"

cls
define res4 ##
res4=integer(micro4)*10
label res4 "Microscopy result of 4th examination"

labelvalue res1-res4 / 0="Negative"
labelvalue res1-res4 /10="1+ positive"
labelvalue res1-res4 /10="1+ positive"
labelvalue res1-res4 /20="2+ positive"
labelvalue res1-res4 /30="3+ positive"
labelvalue res1-res4 /90="No res recorded"
labelvalue res1-res4 /40="Positive, not quantified"
labelvalue res1-res4 /50="Scanty, not quantified"
labelvalue res1-res4 / 1="Scanty, 1 AFB per 100 fields"
labelvalue res1-res4 / 2="Scanty, 2 AFB per 100 fields"
labelvalue res1-res4 / 3="Scanty, 3 AFB per 100 fields"
labelvalue res1-res4 / 4="Scanty, 4 AFB per 100 fields"
labelvalue res1-res4 / 5="Scanty, 5 AFB per 100 fields"
labelvalue res1-res4 / 6="Scanty, 6 AFB per 100 fields"
labelvalue res1-res4 / 7="Scanty, 7 AFB per 100 fields"
labelvalue res1-res4 / 8="Scanty, 8 AFB per 100 fields"
labelvalue res1-res4 / 9="Scanty, 9 AFB per 100 fields"

cls
label sputum1 "Macroscopic sputum of 1st examination"
label sputum2 "Macroscopic sputum of 2nd examination"
label sputum3 "Macroscopic sputum of 3rd examination"
label sputum4 "Macroscopic sputum of 4th examination"
labelvalue sputum1-sputum4 /1="Mucoid"
labelvalue sputum1-sputum4 /2="Purulent"
labelvalue sputum1-sputum4 /3="Muco-purulent"
labelvalue sputum1-sputum4 /4="Blood-tinged"

```

```

labelvalue sputum1-sputum4 /5="Salivary"
labelvalue sputum1-sputum4 /6="Other"
labelvalue sputum1-sputum4 /9="Unknown"

cls
define res1c _
                res1c="-"
if res1= 0      then res1c="N"
if res1> 0 and res1<90 then res1c="P"
if res1=90     then res1c="9"

cls
define res2c _
                res2c="-"
if res2= 0      then res2c="N"
if res2> 0 and res2<90 then res2c="P"
if res2=90     then res2c="9"

cls
define res3c _
                res3c="-"
if res3= 0      then res3c="N"
if res3> 0 and res3<90 then res3c="P"
if res3=90     then res3c="9"

cls
define res4c _
                res4c="-"
if res4= 0      then res4c="N"
if res4> 0 and res4<90 then res4c="P"
if res4=90     then res4c="9"

define pattern ____
pattern=res1c+res2c+res3c+res4c
label pattern "Pattern of 4 serial smears"

* freq pattern

cls
define case #
                case=0
if substr(pattern,1,1)="P" then case=1
if substr(pattern,2,1)="P" then case=1
if substr(pattern,3,1)="P" then case=1
if substr(pattern,4,1)="P" then case=1

label case "Case definition by microscopy"
labelvalue case /0="Negative"
labelvalue case /1="Positive"

define yield ____
if substr(pattern,1,3)="NNN" then yield="NNN"
if substr(pattern,1,3)="NN9" then yield="NN9"
if substr(pattern,1,3)="N99" then yield="N99"
if substr(pattern,1,3)="NN-" then yield="NN9"
if substr(pattern,1,3)="N--" then yield="N99"
if substr(pattern,1,1)="P" then yield="Px"
if substr(pattern,1,2)="NP" then yield="NPx"
if substr(pattern,1,3)="NNP" then yield="NNP"
if substr(pattern,1,4)="NN9P" then yield="NNP"
label yield "Incremental yield of first 3 smears"

keep patid sex marital \
    dateexam1 dateexam2 dateexam3 dateexam4 \
    sputum1 sputum2 sputum3 sputum4 \
    res1 res2 res3 res4 \

```

```

        case pattern yield
savedata "d_ex01_examinee.rec" /replace

*****
* Produce tables on smear pattern and incremental yield

cls
close
read "d_ex01_aggregate.rec"

set option graph /sizex=400
set graph footnote="d_ex01_aggregate"

set echo=off
boxplot meabs /by=sex /bw /sub="    Female                Male"

close
read "d_ex01_examinee.rec"

title "Table 1.  Pattern of serial smear results"
tables case pattern
select case=1
title "Table 2.  Incremental yield among positive results"
tables sex yield /c /PCT
set echo=on

*****
define yesno # global
set echo=off
yesno=?Delete graphs: 1=yes 0=no?
imif yesno=1 then
    erasepng /all /noconfirm
    select
    cls
    type "All graphs erased" /h2
else
    select
endif
set echo=on

*****
* Clean up
set echo=off
yesno=?Delete temporary files: 1=yes 0=no?
imif yesno=1 then
    erase "templ.rec"
    erase "templ.chk"
    erase "temp2.rec"
    erase "temp2.chk"
    select
    cls
    type "All temporary files erased" /h2
    type "File D_EX01_EXAMINEE.REC remains open" /h2
else
    select
endif
set echo=on

```