

Heights – Older Cohort

General heights calculation – see younger and mid-aged section of data dictionary supplement:

- a) Height reported only in cm as required
- b) Height reported in feet and inches will be converted to cm
- c) Height reported in inches only will be converted to cm
- d) All height variables are missing, height is missing

In Surveys 1-3 extreme values (>190cm, <120cm) were deleted. In 2006 it was decided that extreme values should be audited and unless a data-entry error was identified, the value should stand. This applies to past and future surveys.

The literature and our data suggest that women in the older age lose height. The Baltimore Longitudinal Study of Ageing measured women five times over a period of 15 years and found that a cumulative height loss from age 30 to 70 years averaged about 5cm and that this increases to 8cm by the age of 80years (not self-reported data) (Sorkin et al). In more detail it suggests a 95%CI of [-1.024; 0.262] cm for annual changes in height for women aged 70-89. This would allow a loss of approximately 9cm and a growth of 3cm between S1 and S4 for our cohort.

Based on our data and findings of the Baltimore Longitudinal Study of Ageing the following changes in heights between surveys were allowed.

Table 1 Maximum weight change between surveys.

Between surveys	max. height loss	max. height gain
S4S1*	-9	+3
S4S2	-6	+3
S4S3	-5	+3
S3S1	-6	+3
S3S2	-5	+3
S2S1	-5	+3

*S4S1=height in Survey4-height in Survey1, S4S2=height in Survey4-height in Survey2 etc.

Correcting extreme values for height in Survey 4

At Survey 4 there were 50 women with a height >190cm and 74 women with a height <120cm (=124 women with extreme values for height).

In order to identify wrong extreme values in heights for the Older cohort in Survey 4, raw data (which includes extreme values) for Survey 1, Survey 2 and Survey 3 were used to observe changes in heights over time. If for example a women's height was around 115cm in all four surveys, the height was considered to be correct and was therefore not deleted. The following checks were undertaken to determine whether an extreme value was acceptable or wrong.

At first the differences between heights in Survey 4 and heights in the previous surveys were calculated and when the changes were within limits an indicator variable was set to 0, if not to 1. Note, these checks have only been conducted for the 124 women who had extreme values in Survey 4.

```
diff41=ht4cm-ht1cm; IF -1000<diff41<-9 OR diff41>3 THEN check41=1; ELSE  
check41=0;  
diff42=ht4cm-ht2cm; IF -1000<diff42<-6 OR diff42>3 THEN check42=1; ELSE  
check42=0;  
diff43=ht4cm-ht3cm; IF -1000<diff43<-5 OR diff43>3 THEN check43=1; ELSE  
check43=0;
```

Using the indicator variables above and the number of missing values in previous surveys, extreme heights (>190cm;<120cm) for Survey 4 were defined as wrong for the combinations in the Table below.

Number of missing values in Surveys 1 to 3	Number of differences (between Survey 4 and previous surveys) outside the limit
0 or 1	>1
2	>0
3	N/A

This means that if the number of missing values in previous surveys was 0 or 1 and more than one difference was wrong, then the value for Survey 4 was considered as wrong. If the number of missing values in previous surveys was 2 and any of the differences were wrong, then the values for Survey 4 were considered as wrong as well. If the number of missing values was 3 and in all other cases then the value for Survey 4 was considered to be true (OK).

```
IF nmiss (of hts{*}) IN (0,1) AND SUM(check41, check42, check43)>1 THEN
heightchange='WRONG';

ELSE IF nmiss (of hts{*})=2 AND SUM(check41, check42, check43)>0 THEN
heightchange='WRONG';

ELSE IF nmiss (of hts{*})=3 THEN heightchange='OK';

ELSE heightchange='OK';
```

Those heights for Survey 4 which were defined as wrong, were investigated further, to find out whether the height in centimetres was accidentally entered into the foot box or feet were entered into the inch box or etc. A summary is found in the following table.

Table 2 Summary of possible mistakes, their interpretation, proportion of heights recovered.

Rule	cms	feet	inches	Possible interpretations	heights recovered
1	.	. or 0	X(=any value)	Feet were entered into the inch box. newfeet=X;newinches=0. The new value for Survey 4 is then compared with the previous surveys as described above, and defined as OK or wrong. If still wrong, height in Survey 4 is set to missing.	3 of 9
2	.	X	>=12	eg. 5foot 50inches a) 5foot 5inches b) 5foot 0inches Both possibilities were compared with previous surveys and the one that was closest to the mean of the previous surveys, but not the maximum of all four surveys was considered as the new value for Survey 4.	27 of 32
3	.	X	<12	eg. 6foot 9inches->5foot 9inches 3foot 5inches->5foot 5inches It was assumed that a wrong number for feet has been entered and was set to 5. The new value for Survey 4 was then compared with the previous surveys as described above, and defined as OK or wrong. If still wrong height in Survey 4 was set to missing.	15 of 16
4	X	X	X	a)use feet and inches instead of cms b)do cms+100 Both possibilities were compared with previous surveys and the one that was closest to the mean of the previous surveys, but not the maximum of all four surveys was considered as the new value for O4.	16 of 17
5	X	.	.	a) do cms+100 b) do cms*2.54 c) do cms*10 d) eg 58 -> 5foot8inches All four possibilities were compared with previous surveys and the one that was closest to the mean of the previous surveys, but not the maximum of all four surveys was considered as the new value for Survey 4.	29 of 50

With this, 90 out of 124 extreme values for height in Survey 4 were recovered. The height in Survey 4 was set to missing for the remaining 36 women.

Comparing heights over time and filling in missing values/substitute wrong values in all surveys using least square fit

The changes in height over time were observed for all women in the older cohort. A big challenge was the number of missing values per woman across surveys and in which survey(s) they occurred. If any of the following conditions were fulfilled (i.e. at least one height change was strange), then there was 'something wrong' with the heights and heights for these women were checked and the following actions were taken.

IF numbermiss=0 AND (-9>diff41 OR diff41>3 OR -6>diff42 OR diff42>3 OR -5>diff43 OR diff43>3 OR -6>diff31 OR diff31>3 OR -5>diff32 OR diff32>3 OR -5>diff21 OR diff21>3)

1. no missing values

o1htcm< o4htcm	o1htcm> o4htcm	o4htm- o1htcm	o2htcm and o3htcm between o1htcm and o4htcm	Action
✓	-	>3	✓	Problem: growing overall more than allowed. Look for the smallest difference between two consecutive surveys that is smaller than 3. Take the average of these two surveys and fill in all four surveys. (N=128)
✓	-	<=3	✓	Take the mean of all four surveys for each survey. (N=0) (only mentioned for completion, am happy to delete)
-	✓	>=-9	✓	Overall the change over time is correct, and therefore all values will be accepted. (N=219)
-	✓	<-9	✓	Problem: shrinking over time more than allowed. Therefore it is assumed either height in Survey 1 or Survey 4 is wrong. If the absolute difference between Survey 1 and Survey 2 is smaller than the absolute difference between Survey 3 and Survey 4 then height in Survey 4 is set to missing and vice versa. (N=227)

Records with no missing values which did not fit into any of the categories listed in the table above, were dealt with in the following way. In order to find out which one of the four surveys had a wrong value, mean values of absolute differences of three surveys were calculated. In the program they were referred to as blocks, eg. Block123=mean(of abs(diff21) abs(diff32) abs(diff31)). The block that had the smallest mean indicates least height change and if the difference of the first and the last survey in this block were also within limits, then the survey NOT included in the block is set to missing. For example, the following heights were reported by a woman in the older cohort: S1=150cm, S2=145cm, S3=150cm and S4=147cm. Block123=3.33, block234=3.33, block124=3.33, block134=2. Block134 has the smallest mean value of absolute differences (between S3S1, S4S3 and S4S1) and therefore the height in S2 (not included in block134) was set to missing (after filling in missing values, this women ended up with 150cm in Survey2). This method was applied to N=1100 women.

For any other cases with no missing values, if there were identical heights at two surveys, the height was set to the value that appeared twice (N=61).

2. one missing value

If one is missing and the other values suggest a constant growth

eg. o1htcm=. AND o2htcm<=o3htcm<=o4htcm eg. o2htcm=. AND o1htcm<=o3htcm<=o4htcm	Problem: the women are growing (N=1235). The average of those two heights with the least difference <=3 will be used to replace height in all surveys (N=69). For women with the same height in two or three surveys, the missing height will be set to the same height (N=1166).
--	--

For other cases with one missing value, the following four tables show whether differences in height between surveys are within the limit (OK) or outside the limit (NO), also considering in which survey the missing value occurred.

O1htcm=.

S2S3	S3S4	S2S4	Action	N
OK	OK	OK	No action	140
OK	OK	NO	Survey 4 is wrong, set to missing.	7
OK	NO	OK	IF o4htcm<=o2htcm THEN Survey 3 is wrong, else Survey 4 is wrong, set to missing.	15
OK	NO	NO	Survey 4 is wrong, set to missing.	7
NO	OK	OK	If o4htcm<=o2htcm then Survey 3 is wrong, else Survey 2 is wrong, set to missing.	21
NO	OK	NO	Survey 2 is wrong, set to missing.	16
NO	NO	OK	Survey 3 is wrong, set to missing.	20
NO	NO	NO	All are wrong, set to missing.	5

O2htcm=.

S1S3	S3S4	S1S4	Action	N
OK	OK	OK	No action	323
OK	OK	NO	Survey 4 is wrong, set to missing.	1
OK	NO	OK	IF o4htcm<=o1htcm then S3 is wrong, else Survey 4 is wrong, set to missing.	38
OK	NO	NO	Survey 4 is wrong, set to missing.	19
NO	OK	OK	IF o4htcm<=o1htcm then Survey 3 is wrong, else Survey 1 is wrong, set to missing.	18
NO	OK	NO	Survey 1 is wrong, set to missing.	12
NO	NO	OK	Survey 3 is wrong, set to missing.	24
NO	NO	NO	All wrong, set to missing.	7

O3htcm=.

S1S2	S2S4	S1S4	Action	N
OK	OK	OK	No action	402
OK	OK	NO	Survey 4 is wrong, set to missing.	5
OK	NO	OK	If o4htcm<=o1htcm then Survey 2 is wrong, else Survey 4 is wrong, set to missing.	36
OK	NO	NO	Survey 4 is wrong, set to missing.	20
NO	OK	OK	If o4htcm<=o1htcm then Survey 2 is wrong, else Survey 1 is wrong, set to missing.	39
NO	OK	NO	Survey 1 is wrong, set to missing.	8
NO	NO	OK	Survey 2 is wrong, set to missing.	48
NO	NO	NO	All wrong, set to missing.	6

O4htcm=.

S1S2	S2S3	S1S3	Action	N
OK	OK	OK	No action	874
OK	OK	NO	Survey 3 is wrong, set to missing.	29
OK	NO	OK	If o3htcm<=o1htcm then Survey 2 is wrong, else Survey 3 is wrong, set to missing.	65
OK	NO	NO	Survey 3 is wrong, set to missing.	135
NO	OK	OK	If o3htcm<=o1htcm then Survey 2 is wrong, else Survey 1 is wrong, set to missing.	60
NO	OK	NO	Survey 1 is wrong, set to missing.	48
NO	NO	OK	Survey 2 is wrong, set to missing.	122
NO	NO	NO	All wrong, set to missing.	25

3. two missing values (n=1146)

Condition	Action
IF the difference between the two heights is between 0 and 3	Use the mean of the two heights to fill in the missing values
IF the difference between the two heights is not between the limits	Set all to missing

4. more than two missing values

Condition	Action
If all are missing	All stay missing
If one is not missing	Keep that value if it is between 120 and 190.

After deciding which values were right or wrong, the next step was to fill in missing values where there were one or two missing values in all four surveys, using the least square fit. Unfortunately I wasn't able to find a suitable macro or a function which would have allowed me to do it SAS, therefore I had to use the Trend() function in Excel. There were N=1146 records with two missing values filled in and N=1739 records with a single missing value. NOTE: at this stage heights were created for women who weren't in the study any more (withdrawals or deaths).

After checking that the new values were within the limits, 151 of the 1146 double missing values had to be set back to missing. The main reason for this was that both values were for adjacent surveys and already 5cm apart. The following table shows statistics for the old height variable (o1htcmold) and the new one (newo1htcm). (These tables only include women who are still in the study at the survey, eg there is no height for Survey 3 and Survey 4 if the women dropped out after Survey 2)

Table 3 Descriptive statistics for the old and new height variable for all four surveys.

Variable	N		25th		75th		Max	Mean
	N	Miss	Min	Pctl	Median	Pctl		
<i>o1htcmold</i>	12362	578	120.0	157.0	160.0	165.0	188.0	161.2
<i>newo1htcm</i>	12051	889	124.0	157.0	161.0	165.0	188.0	161.4
<i>o2htcmold</i>	9704	3236	120.0	156.0	160.0	165.0	188.0	160.5
<i>newo2htcm</i>	9956	2984	129.0	157.0	160.0	165.0	185.0	160.7
<i>o3htcmold</i>	7987	4953	122.0	155.0	160.0	165.0	188.0	159.7
<i>newo3htcm</i>	8478	4462	127.0	155.0	160.0	165.0	185.0	160.2
<i>newo4htcm</i>	6885	6055	127.0	155.0	160.0	163.0	185.0	159.6

Although the number of missing values increased in Survey 1, all subsequent surveys had fewer missing values. The overall statistics do not change.

Table 4 Differences between surveys BEFORE data cleaning.

Variable	N	Miss	25th Min	Pctl	75th Median	Pctl	Max	Mean
S4S1old	6388	6552	-163.0	-4.000	-1.000	0	510.0	-2.444
S4S2old	6045	6895	-168.0	-3.000	0	0	383.0	-1.821
S4S3old	5964	6976	-165.0	-2.000	0	1.000	512.0	-0.760
S2S1old	9354	3586	-41.00	-2.000	0	0	33.00	-0.753
S3S1old	7732	5208	-41.00	-3.000	0	0	53.00	-1.680
S3S2old	7314	5626	-43.00	-2.000	0	0	51.00	-0.872

Table 5 Differences between surveys AFTER data cleaning.

Variable	N	N Miss	Min	25th Pctl	Median	75th Pctl	Max	Mean
S4S1	6861	6079	-9.000	-3.000	0	0	3.000	-1.849
S4S2	6861	6079	-9.000	-3.000	0	0	3.000	-1.140
S4S3	6872	6068	-9.000	-2.000	0	0	3.000	-0.582
S2S1	9818	3122	-8.000	-2.000	0	0	3.000	-0.753
S3S1	8428	4512	-9.000	-3.000	0	0	3.000	-1.290
S3S2	8450	4490	-8.000	-2.000	0	0	8.000 *	-0.596

*although the rule was not more than 3cm plus there are a couple for which it exceeds this limit. These cases fall mainly into the category no missing values rule no 3, which says if S4 is smaller than S1(no more than 9cm smaller) and S2 and S3 are between S1 and S4 then all values are being kept. This particular case has the height sequence of 173, 165, 173, 165.

Table 4 and Table 5 show nicely that the cleaning had a big effect on the data in terms of 'making sense'.