Correction of health index data

A.J.Nicholson

Introduction

This documents record corrections made to the Health Index data in light of the issues discovered in "ExploratoryAnalysis.rmd".

```
cols<-read.table("data/cols.txt",colClasses="character");cols<-as.character(cols[2,])
HIdata_new<-read.table("data/HIdata.txt",header=TRUE,colClasses=cols)
Issues<-read.table("data/Issues.txt",header=TRUE)
levels(Issues$issue)
```

```
##  [1] "33 sockets"                     "Abscesses should be NA"
##  [3] "age outside age range"          "Cavities but no teeth"
##  [5] "Erroneous Cavities"             "Max and Min but no Age"
##  [7] "MaxAge lower than MinAge"       "Min and Max age should be NA"
##  [9] "No sockets but Abscesses"       "Non-juvenile deciduous teeth"
## [11] "Non Juvenile femoral diaphysis" "Not sex 5"
## [13] "Single digit social code"       "SUMPRE miscounted"
## [15] "SUMTET+SUMPRE >32"              "Too many teeth"
## [17] "too young to sex"
```

In addition to these the problem of NA values in the categorical variables will be addressed.

1. Not sex 5

These are individuals categorized as sex group 5 (Sex is undetermined because juvenile) that aren't juvenile. it Will correct by changing ID to 6 (Unknown).

```
IDs<-Issues$ID[Issues$issue=="Not sex 5"]
HIdata_new$Sex[HIdata_new$cID %in% IDs]<-"6"
```

2. Too young to sex

Juveniles, according to the age range, but have be sexed. Correct by putting in to sex class 5.

```
IDs<-Issues$ID[Issues$issue=="too young to sex"]
HIdata_new$Sex[HIdata_new$cID %in% IDs]<-"5"
```

3. MaxAge lower than MinAge

This is where the max and min are the wrong way around. Fixed by switching values

```
IDs<-Issues$ID[Issues$issue=="MaxAge lower than MinAge"]
HIdata_new[HIdata_new$cID %in% IDs,c("MinAge","MaxAge")]<-HIdata_new[HIdata_new$cID %in% IDs,c("MaxAge","MinA
```

4. Max and Min but no Age

Minimum (MinAge) and maximum (MaxAge) ages are recorded but not Age. I Will use the average of min and max as an approximation of Age

```
IDs<-Issues$ID[Issues$issue=="Max and Min but no Age"]
HIdata_new$Age[HIdata_new$cID %in% IDs]<-(HIdata_new$MaxAge[HIdata_new$cID %in% IDs]+HIdata_new$MinAge[HIdata
```

5. Age outside age range

The problem is that the age given is outside the age range (MinAge to MaxAge). Corrected by replacing Age with the average of MinAge and MaxAge.

```
IDs<-Issues$ID[Issues$issue=="age outside age range"]
HIdata_new$Age[HIdata_new$cID %in% IDs]<-(HIdata_new$MaxAge[HIdata_new$cID %in% IDs]+HIdata_new$MinAge[HIdata
```

6. Min and Max age should be NA

These are cases where the min and max are recorded as 0 as a substitute for missing values. Fixed by replacing with NA

```
IDs<-Issues$ID[Issues$issue=="Min and Max age should be NA"]
HIdata_new[HIdata_new$cID %in% IDs,c("MinAge","MaxAge")]<-NA
```

7.:Non Juvenile femoral diaphysis

Problem: femoral diaphysis length is growth measure in juveniles where the epiphysis haven't full fused, can't record for adult remains. Solved by replacing the value in FDIAP with NA

```
IDs<-Issues$ID[Issues$issue=="Non Juvenile femoral diaphysis"]
HIdata_new$FDIAP[HIdata_new$cID %in% IDs]<-NA
```

9. SUMTET+SUMPRE >32

This is where the number of teeth lost antemortem (so can't have been recovered) plus the total teeth recorded is more than 32. In these case the difference is so great (there are far too many teeth) that recalculating one or other value isn't a reasonable approach; so the issue is fixed by changing SUMPRE to NA

```
IDs<-Issues$ID[Issues$issue=="SUMTET+SUMPRE >32"]
HIdata_new$SUMPRE[HIdata_new$cID %in% IDs]<-NA
```

10. Cavities but no teeth

This is where cavities have been recorded but the total number of teeth observed hasn't (i.e. no SUMTET). This is corrected by estimating SUMTET to be 32-SUMPRE

```
IDs<-Issues$ID[Issues$issue=="Cavities but no teeth"]
HIdata_new$SUMTET[HIdata_new$cID %in% IDs]<-32-HIdata_new$SUMPRE[HIdata_new$cID %in% IDs]
```

11. 33 sockets

33 sockets have been recorded (in SUMSOK) instead of 32. Corrected by replacing 33 with 32.

```
IDs<-Issues$ID[Issues$issue=="33 sockets"]
HIdata_new$SUMSOK[HIdata_new$cID %in% IDs]<-32
```

12. Single digit social code

The social code (SOC) should be a three digit number but is recorded as "1" in a number of records. This is fixed by replacing 1 with 111 (code for an undifferentiated society or one with no evidence of differentiation)

```
IDs<-Issues$ID[Issues$issue=="Single digit social code"]
HIdata_new$SOC[HIdata_new$cID %in% IDs]<-111
```

13. Erroneous Cavities

These are cavities recorded without teeth (SUMTET=0 and SUMPRE(antemortem loss)=32).Fixed by changing SUMCAV to NA

```
IDs<-Issues$ID[Issues$issue=="Erroneous Cavities"]
HIdata_new$SUMCAV[HIdata_new$cID %in% IDs]<-NA
```

14. SUMPRE miscounted

This is where the number of teeth lost antemortem plus the total teeth recorded is more than 32, but the difference is small so could be miscounting. solution = changing SUMPRE to 32-SUMTET.

```
IDs<-Issues$ID[Issues$issue=="SUMPRE miscounted"]
HIdata_new$SUMPRE[HIdata_new$cID %in% IDs]<-32-HIdata_new$SUMTET[HIdata_new$cID %in% IDs]
```

15. Abscesses should be NA

This is where Abscesses have been recorded but not the number of sockets observed(SUMSOK) for these there is no way of estimating SUMSOK so the only solution is to change SUMABS to NA.

```
IDs<-Issues$ID[Issues$issue=="Abscesses should be NA"]
HIdata_new$SUMABS[HIdata_new$cID %in% IDs]<-NA
```

16. No sockets but Abscesses

As the previous issue Abscesses have been recorded but not the number of sockets observed(SUMSOK). In this case the SUMSOK can be estimated by taking the number of teeth lost antemortem (SUMPRE) from 32.

```
IDs<-Issues$ID[Issues$issue=="No sockets but Abscesses"]
HIdata_new$SUMSOK[HIdata_new$cID %in% IDs]<-32-HIdata_new$SUMPRE[HIdata_new$cID %in% IDs]
```

17. Non-juvenile deciduous teeth

Linear enamel hypoplasia recorded for Decidous canines from adult remains. Change LDC to 0 to correct.

```
IDs<-Issues$ID[Issues$issue=="Non-juvenile deciduous teeth"]
HIdata_new$LDC[HIdata_new$cID %in% IDs]<-0
```

18. NA value in category variable

For category variable the codebook defines 0 as un-recordable(or similar) so there shouldn't be any NA's. To fix this i will cycle through all these columns and replace NA with 0.

```
#Hypoplasia
HIdata_new$LDI[is.na(HIdata_new$LDI)]<-0;HIdata_new$LDC[is.na(HIdata_new$LDC)]<-0
HIdata_new$LPI[is.na(HIdata_new$LPI)]<-0;HIdata_new$LPC[is.na(HIdata_new$LPC)]<-0
#Anemia
HIdata_new$CROB[is.na(HIdata_new$CROB)]<-0;HIdata_new$PORHY[is.na(HIdata_new$PORHY)]<-0
#Auditory exostosis
HIdata_new$AUDEX[is.na(HIdata_new$AUDEX)]<-0
#Degenerative joint disease
HIdata_new$DJSH[is.na(HIdata_new$DJSH)]<-0;HIdata_new$DJHK[is.na(HIdata_new$DJHK)]<-0
HIdata_new$DJCER[is.na(HIdata_new$DJCER)]<-0;HIdata_new$DJTHO[is.na(HIdata_new$DJTHO)]<-0
HIdata_new$DJLUM[is.na(HIdata_new$DJLUM)]<-0;HIdata_new$DJTMJ[is.na(HIdata_new$DJTMJ)]<-0
HIdata_new$DJWR[is.na(HIdata_new$DJWR)]<-0;HIdata_new$DJHAN[is.na(HIdata_new$DJHAN)]<-0
#Trauma
HIdata_new$TRARM[is.na(HIdata_new$TRARM)]<-0;HIdata_new$TRLEG[is.na(HIdata_new$TRLEG)]<-0
HIdata_new$TRNAS[is.na(HIdata_new$TRNAS)]<-0;HIdata_new$TRARM[is.na(HIdata_new$TRARM)]<-0
HIdata_new$TRFAC[is.na(HIdata_new$TRFAC)]<-0;HIdata_new$TRSKUL[is.na(HIdata_new$TRSKUL)]<-0
HIdata_new$TRHAN[is.na(HIdata_new$TRHAN)]<-0;HIdata_new$TRWEAP[is.na(HIdata_new$TRWEAP)]<-0
#Infection
HIdata_new$TIBINF[is.na(HIdata_new$TIBINF)]<-0
```

The only categorical variable in which zero isn't used to represent unavailable data is SKELINF. For this the coding is a follows.

0. "no periosteal reaction on any other bone than the tibiae"
1. "periosteal reaction on any other bone(s) than the tibiae not caused by trauma"
2. "evidence of systemic infection involving any of the bones (including the tibiae) of the skeleton. This would include specific diseases which include (but are not limited to) tuberculosis and syphilis"

It is odd that this coding doesn't match the general pattern. As such I will change the coding to the following:

0. Unknown/un-recordable
1. "no periosteal reaction on any other bone than the tibiae"
2. "periosteal reaction on any other bone(s) than the tibiae not caused by trauma"
3. "evidence of systemic infection involving any of the bones (including the tibiae) of the skeleton. This would include specific diseases which include (but are not limited to) tuberculosis and syphilis"

```
Fact2Num<- function(x){
# function from HIBasicClean.R
a<-as.character(x)
a<-as.numeric(a)
```

```
a
}

HIdata_new$SKELINF<-Fact2Num(HIdata_new$SKELINF)+1
HIdata_new$SKELINF[is.na(HIdata_new$SKELINF)]<-0
HIdata_new$SKELINF<-as.factor(HIdata_new$SKELINF)
```

Review data

```
kable(summary(HIdata_new[,3:9]))
```

Sex	Age	DentalAge	MinAge	MaxAge	Ancestry	SOC
1 :1943	Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. : 0.00	1:9818	111 :3333
2 :1531	1st Qu.:11.80	1st Qu.: 0.600	1st Qu.:14.00	1st Qu.: 9.00	2:1305	232 :2948
3 :1794	Median :32.00	Median : 2.200	Median :20.00	Median :30.00	3:1377	332 :1050
4 :1365	Mean :26.65	Mean : 4.455	Mean :21.77	Mean :28.78	5: 7	333 : 486
5 :3313	3rd Qu.:36.50	3rd Qu.: 7.000	3rd Qu.:31.00	3rd Qu.:44.00	6: 11	252 : 239
6 :2437	Max. :80.00	Max. :20.000	Max. :70.00	Max. :99.00	NA	(Other): 531
NA's: 135	NA's :706	NA's :9859	NA's :1239	NA's :3362	NA	NA's :3931

```
kable(summary(HIdata_new[,10:16]))
```

FDIAP	FLEN	HEIGT	FMIDA	FMIDM	HLEN	HCIR
Min. : 40.0	Min. :308.0	Min. :1245	Min. :16.00	Min. :15.00	Min. :220.0	Min. :32.00
1st Qu.: 90.0	1st Qu.:406.0	1st Qu.:1549	1st Qu.:26.00	1st Qu.:24.00	1st Qu.:291.0	1st Qu.:57.00
Median :156.0	Median :428.0	Median :1609	Median :28.00	Median :25.00	Median :306.0	Median :62.00
Mean :179.9	Mean :429.0	Mean :1609	Mean :27.81	Mean :25.59	Mean :307.1	Mean :62.41
3rd Qu.:253.2	3rd Qu.:451.2	3rd Qu.:1670	3rd Qu.:30.00	3rd Qu.:27.00	3rd Qu.:323.0	3rd Qu.:68.00
Max. :475.0	Max. :556.0	Max. :1880	Max. :40.00	Max. :47.00	Max. :398.0	Max. :90.00
NA's :11278	NA's :9686	NA's :9469	NA's :9563	NA's :9552	NA's :10182	NA's :9852

```
kable(summary(HIdata_new[,17:25]))
```

LDI	LDC	LPI	LPC	SUMTET	SUMPRE	SUMCAV	SUMSOK	SUMABS
0:11931	0:11892	0:9867	0:8976	Min. : 0.00	Min. : 0.000	Min. : 0.000	Min. : 0.00	Min. :0.0000
1: 556	1: 588	1:1896	1:2025	1st Qu.: 0.00	1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.: 0.00	1st Qu.:0.0000
2: 28	2: 33	2: 484	2: 945	Median : 7.00	Median : 0.000	Median : 0.000	Median : 0.00	Median :0.0000
3: 3	3: 5	3: 271	3: 572	Mean :10.81	Mean : 1.761	Mean : 1.238	Mean :10.09	Mean :0.4609
NA	NA	NA	NA	3rd Qu.:22.00	3rd Qu.: 1.000	3rd Qu.: 1.000	3rd Qu.:21.00	3rd Qu.:0.0000
NA	NA	NA	NA	Max. :38.00	Max. :32.000	Max. :25.000	Max. :32.00	Max. :9.0000
NA	NA	NA	NA	NA's :1916	NA's :2637	NA's :2339	NA's :2097	NA's :2563

```
kable(summary(HIdata_new[,26:37]))
```

CROB	PORHY	AUDEX	TIBINF	SKELINF	DJSH	DJHK	DJCER	DJTHO	DJLUM	DJTMJ	DJWR
0:7263	0:6674	0:9386	0:6551	0: 579	0:8031	0:8064	0:9144	0:9394	0:9263	0:9304	0:9010
1:4270	1:5054	1:2988	1:4099	1:9718	1:3030	1:2918	1:2371	1:2032	1:1905	1:2524	1:2807
2: 955	2: 762	2: 144	2:1286	2:1470	2:1193	2:1239	2: 658	2: 781	2: 695	2: 690	2: 701

CROB	PORHY	AUDEX	TIBINF	SKELINF	DJSH	DJHK	DJCER	DJTHO	DJLUM	DJTMJ	DJWR
3: 30	3: 28	NA	3: 398	3: 751	3: 246	3: 277	3: 276	3: 256	3: 539	NA	NA
NA	NA	NA	4: 184	NA	4: 14	4: 17	4: 69	4: 55	4: 116	NA	NA
NA	NA	NA	NA	NA	5: 4	5: 3	NA	NA	NA	NA	NA

```
kable(summary(HIdata_new[,38:45]))
```

DJHAN	TRARM	TRLEG	TRNAS	TRFAC	TRSKUL	TRHAN	TRWEAP
0:9768	0:7553	0:6972	0:8985	0:8652	0:7041	0:9824	0: 1589
1:2326	1:4777	1:5393	1:3477	1:3787	1:5164	1:2630	1:10739
2: 424	2: 107	2: 97	2: 56	2: 79	2: 313	2: 64	2: 190
NA	3: 52	3: 36	NA	NA	NA	NA	NA
NA	4: 2	4: 8	NA	NA	NA	NA	NA
NA	5: 27	5: 12	NA	NA	NA	NA	NA

missing Sex and age data

The summary above shows NAs in sex and age.

Sex

For sex there are two unknown categories (cat 5= unknown because juvenile, 6= undeterminable). If possible all NA should be put into one of these categories.

```
SexNA<-HIdata_new[is.na(HIdata_new$Sex),]
kable(summary(SexNA[,4:7]))
```

Age	DentalAge	MinAge	MaxAge
Min. : NA	Min. : NA	Min. : NA	Min. : NA
1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA
Median : NA	Median : NA	Median : NA	Median : NA
Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN
3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA
Max. : NA	Max. : NA	Max. : NA	Max. : NA
NA's :135	NA's :135	NA's :135	NA's :135

All NA on Age so can't tell if cat. 5 or 6 (will leave).

Age

Age is needed for the health index calculation so need to see if age can be estimated when missing.

```
#706 Age NA's (need age for health index calculation)
AgeNA<-HIdata_new[is.na(HIdata_new$Age),]
kable(summary(AgeNA[,3:7]))
```

Sex	Age	DentalAge	MinAge	MaxAge
1 : 56	Min. : NA	Min. : NA	Min. : 0.00	Min. : NA

Sex	Age	DentalAge	MinAge	MaxAge
2 : 89	1st Qu.: NA	1st Qu.: NA	1st Qu.:30.00	1st Qu.: NA
3 : 42	Median : NA	Median : NA	Median :30.00	Median : NA
4 : 76	Mean :NaN	Mean :NaN	Mean :26.29	Mean :NaN
5 : 15	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.:30.00	3rd Qu.: NA
6 :293	Max. : NA	Max. : NA	Max. :40.00	Max. : NA
NA's:135	NA's :706	NA's :706	NA's :689	NA's :706

There are 706 records without ages. None of these have MaxAge or DentalAge recorded but some do have MinAge. In the absence of anything better this could be a substitute for age.

```
Min2Age<-AgeNA[!(is.na(AgeNA$MinAge)),]
kable(Min2Age[,2:7])
```

	ID	Sex	Age	DentalAge	MinAge	MaxAge
5279	37	3	NA	NA	30	NA
5381	032	6	NA	NA	30	NA
5384	1C3	3	NA	NA	30	NA
5387	1C6	3	NA	NA	30	NA
5400	7C4	1	NA	NA	30	NA
5417	17C2	3	NA	NA	30	NA
5423	18C3	3	NA	NA	30	NA
5441	23C10	5	NA	NA	1	NA
5461	27C11	3	NA	NA	30	NA
5464	27C14	6	NA	NA	30	NA
5477	28C10	1	NA	NA	30	NA
5487	28C24	2	NA	NA	30	NA
5489	28C27	5	NA	NA	0	NA
5491	28C29	3	NA	NA	40	NA
5534	3	1	NA	NA	30	NA
8702	2942	6	NA	NA	25	NA
8730	2970	6	NA	NA	21	NA

17 observations, all juvenile MinAge are cat 5 in sex, and the "adult" are not all 18+ (i.e. not just we know it adult but nothing better). so MinAge can be used as a sub for Age so these observations are useable.

```
IDs<-Min2Age$cID
HIdata_new$Age[HIdata_new$cID %in% IDs]<-HIdata_new$MinAge[HIdata_new$cID %in% IDs]
```

Next step

This concludes the correction of the Health Index data. The next stage of analysis (detailed in "HealthIndex.rmd") is reconstruction of process for creating the health index.

```
write.table(HIdata_new,"data/HIdata_c.txt",row.name=FALSE)
```