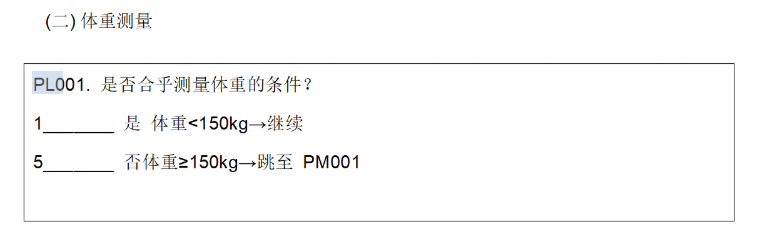
**表1 – 体检数据缺失值**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **变量名** | **原名** | **变量含义** | **变量值** | **比例/%** |
| num\_rows\_biodata | ------ | 体检数据总行数 | 13974 | 100 |
| num\_height\_mis | qi002 | 身高值缺失 | 265 | 1.90 |
| num\_WeightJudge\_mis | pl001 | 体重判断项缺失(注释见图1) | 91 | 0.65 |
| num\_weight\_mis | ql002 | 体重值缺失 | 246 | 1.76 |
| num\_waist\_mis | qm002 | 腰围值缺失 | 203 | 1.45 |

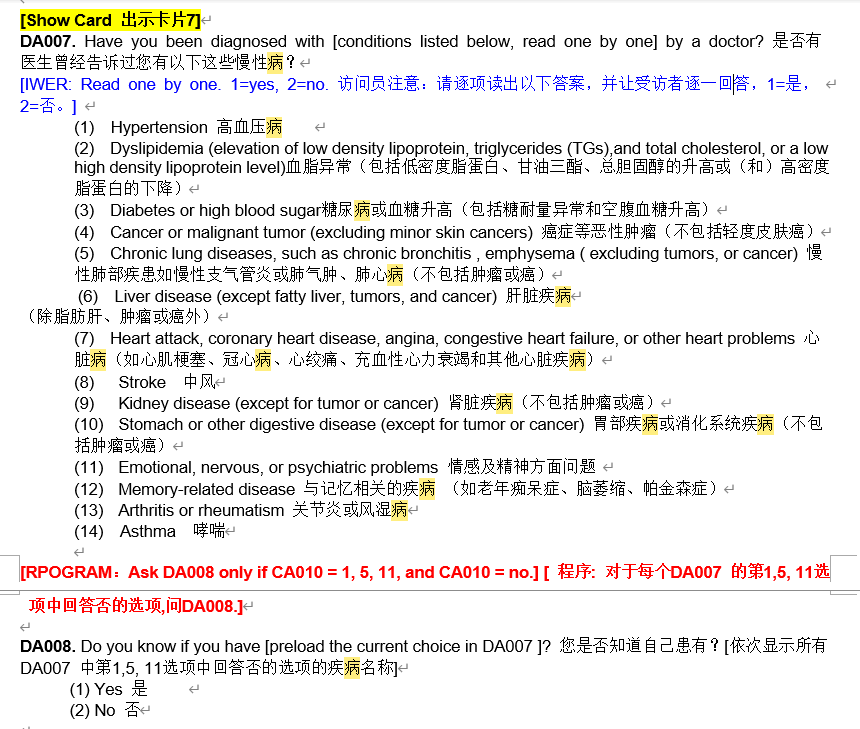
**图1**



**表2 – 健康状况数据缺失值**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **变量名** | **原名** | **变量含义** | **变量值** | **比例/%** |
| num\_rows\_healthdata | ------ | 健康状况数据总行数 | 17596 | 100 |
| num\_hyperten\_mis | da007\_1\_ | 高血压缺失 | 126 | 0.72 |
| num\_hypertenconfirm\_mis | da008\_1\_ | 高血压确认缺失(注释见图2) | 4324 | 24.57 |
| num\_sugar\_mis | da007\_3\_ | 糖血糖尿缺失 | 188 | 1.07 |
| num\_heartatt\_mis | da007\_7\_ | 心脏病缺失 | 129 | 0.73 |
| num\_kidneydis\_mis | da007\_9\_ | 肾病缺失 | 144 | 0.82 |

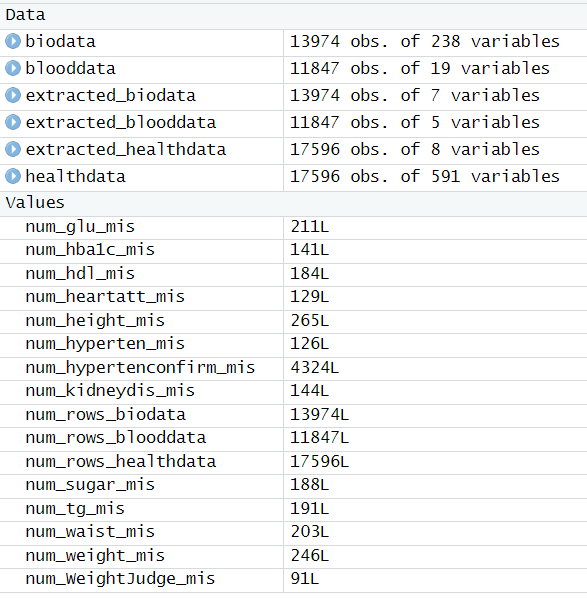
**图2**

****

**表3 – 血检数据缺失值**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **变量名** | **原名** | **变量含义** | **变量值** | **比例/%** |
| num\_rows\_blooddata | ------ | 血检数据总行数 | 11847 | 100 |
| num\_hba1c\_mis | newhba1c | HbA1c缺失 | 141 | 1.19 |
| num\_hdl\_mis | newhdl | 高密度胆固醇缺失 | 184 | 1.55 |
| num\_tg\_mis | newtg | 甘油三酯缺失 | 191 | 1.61 |
| num\_glu\_mis | newglu | 血糖缺失 | 211 | 1.78 |

**图3 – Rstudio中变量对照**



**以下为正文**

此为charls2011基线调查的数据初步整理，包括体检信息(biomarkers.dta)、健康状况与功能(health\_status\_and\_functioning.dta)、血检数据(Blood\_20140429.dta)。将遇到的问题进行简单的总结汇报。

1. 参考表1、表2、表3，有部分缺失值，不知道如何处理。是否在包含缺失值的情况下可根据其它特征进行CKM的判断？ 是部分缺失还是全部缺失，部分缺失进行插补，全部缺失要删除
2. 参考表1、图1，大于等于150KG的个体在体重栏应表现为缺失，是否应赋值或做其他处理？ 离散值标为缺失
3. 参考表2，部分项目（如糖尿糖血等）在问卷中并未体现具体病症名称，仅为笼统概括，而CKM的临床定义包括具体病症名称 结合数据库做适应性调整
4. 参考表2、图2，问卷问题DA008，据我理解应是“医生未告知某人患有A病，而某人自知患有A病”，在数据分析上应按照医生所述还是个人所述？个人所述
5. 参考表3，血糖一项在数据手册中似乎并未重点提出为空腹血糖 结合数据库描述清楚即可
6. 不同数据集中样本数量不同，是否要筛选出仅在三个数据集中均存在的样本？后续继续处理2015年的数据，是否要筛选出仅在2011和2015所有数据集中存在的样本？

目前问题大致如上