· 计算机应用 ·

SAS 软件在医学研究数据核查中的应用*

国家心血管病中心 中国医学科学院 北京协和医学院 阜外医院 国家心血管疾病临床医学研究中心 心血管国家重点实验室(100037)
白雪珂　张小艳　路甲鹏　李　希　吴超群△

【提　要】 目的　使用 SAS 以及操作系统的“任务计划”功能实现医学研究数据实时核查。**方法**　通过 SAS 直接访问数据集,对制定好的逻辑核查类型编写宏程序。**结果**　采用自行编制的 SAS 宏程序及其调用实例、SAS 设置邮件发送和操作系统自动化运行 SAS,完成缺失值以及异常值的核查。**结论**　当前研究所提供的 SAS 宏程序和系统设置,可帮助医学研究实现简单的数据核查功能,保证研究数据质量。

【关键词】 SAS 宏程序　数据核查　定时任务
【中图分类号】 R195.1　　**【文献标识码】** A　　**DOI** 10.11783/j.issn.1002-3674.2025.01.027

随着技术的发展,越来越多的医学研究使用电子数据采集系统(electronic data capture system,EDC)进行数据管理[1]。EDC 一个常见的重要功能是逻辑核查,即数据管理员利用系统嵌入的工具,将预先设计的逻辑规则配置到系统中;当有异常数据触发这些规则时,系统发布质疑(query),提醒研究者进行核对和修正。这类在数据收集阶段即对质量进行实时检查的工作方式,对于保证数据质量具有重要意义。然而,在实际工作中,有一些研究仍无法实现使用 EDC 系统;或仅采用了电子化的数据采集功能(如购买或自行研发电子问卷),但未开发或使用数据核查功能。本文主要针对这一问题,提出利用 SAS 软件、计算机操作系统自带的定时任务功能,实现定时数据核查的功能。

对象与方法

1. 对象

以某心血管疾病领域开展的前瞻性研究为例。该研究的数据采集使用自行研发的电子问卷系统,支持数据在线录入,并将数据实时上传至服务器。自研系统未开发 query 功能。

2. 方法

利用 SAS/ACCESS 模块功能,直接访问据集。然后,基于简单的 SAS 宏程序,对已确定的逻辑核查内容进行统计和报告,并自动发送核查邮件。

(1)访问数据库。研究数据存储在 Oracle 数据库服务器中。数据库管理员在本地 Oracle 客户端进行配置后,利用 SAS/ACCESS 模块的功能,即可通过 SAS 直接访问研究数据。还可进一步通过配置操作系统中的 ODBC 接口,实现更为简洁的数据访问。另外,对于使用 excel 表的研究者可以使用 PROC IMPORT 语法导入 excel 表格。

(2)实施逻辑核查。该前瞻性研究的数据核查主要包括数据缺失、重复值、超出规定的赋值范围、与其他变量的逻辑关系异常、多选题选项互斥、日期范围异常等。针对各个核查类型撰写相应的 SAS 宏程序,可以较为简洁的实现数据核查任务,并生成异常数据报表,供研究者查看。

(3)数据报表的保存和核查邮件发送。可将数据报表存入指定的文件夹地址,供数据管理员和研究者查看。同时利用 SAS 发送电子邮件的功能,在汇总核查数据后以邮件形式自动发送给指定的收件人。

(4)功能自动化。利用操作系统的“任务计划”功能,通过编写 bat 脚本,可以实现 SAS 程序的定时运行。

结　果

1. 访问数据库

在安装最新版本的 Oracle 之后,通过 Oracle Net Manager → 编辑 → 创建,配置对应的数据库地址、端口等,如图 1 所示。

图 1　Oracle 连接数据库

* 基金项目:国家重点研发计划(2023YFC2509400);中国医学科学院医学与健康科技创新工程(CIFMS, 2021-I2M-1-007);中国医学科学院阜外医院高水平医院临床科研业务费(零余额 2022-GSP-GG-4,零余额 2023-GSP-RC-20)
△ 通信作者:吴超群,E-mail:wuchaoqun@ fuwai. com

通过控制面板搜索 ODBC，选择 ODBC 数据管理程序 → 添加 → Oracle → 配置对应的设置，注意 TNS Service Name 为上一步 Oracle 的服务名，如图 2 所示。

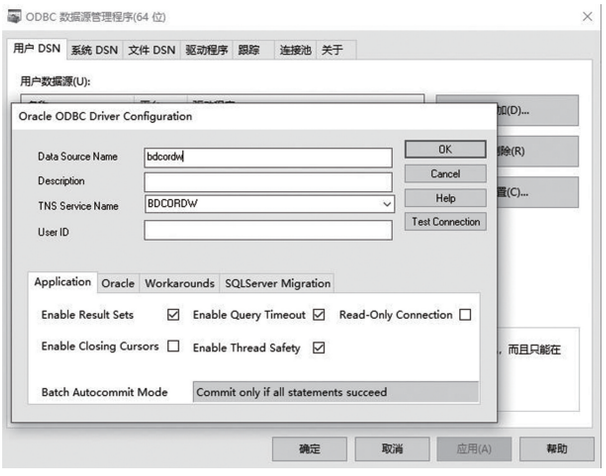

图 2　ODBC 连接数据库

在测试配置成功后，使用如下代码，SAS 即可通过 ODBC 接口调用数据源：

LIBNAME 自定义名称 *ODBC DATASRC* = *ODBC* 配置名称 *USER* = 用户名 *PASSWORD* = 密码；

使用 PROC IMPORT 语法导入 excel 表代码如下：

PROC IMPORT OUT = 输出的 *SAS* 数据集名称 *DATAFILE* = "文件路径、文件名以及扩展名" *DBMS* = *EXCEL REPLACE*；*RUN*；

2. 逻辑核查和结果导出

在逻辑核查时，首先需要数据管理员预先对研究需要核查的内容进行分类整理，概括主要的核查类型，并形成文档记录（表 1）。然后，针对每种类型，撰写适用于该类型数据核查的宏程序。最后，调用宏程序完成数据核查。以下是针对变量缺失和数值变量异常值（超出规定的上下界值）撰写的宏程序，注释以及调用实例。

表 1　核查记录文档示例

核查编号	待核查变量名称	核查名称	核查分类	核查逻辑
001	BRTHDAT	出生日期有缺失	缺失	missing(BRTHDAT)
002	SBP	收缩压异常	异常值	SB*P*<50 mmHg or SBP>280 mmHg
003	DISNONE	和其他选项互斥	多选题互斥	DISNONE='Y' and DIS1='Y' and DIS2='Y'

（1）检查变量缺失的宏程序

```
%MACRO CHECK_MISSING(infile=,var=,varlist=,checkid=,checklabel=, outfile=);
/* 备注：infile 为输入数据，指待核查的目标数据集，可以是在上述访问的原始数据基础上进行处理后的数据集；var 为待核查变量名称；varlist 为导出表格需要保留的变量，通常输出患者 ID，患者所属的项目点名称；checkid 为核查编号，checklabel 为核查名称，outfile 为导出文件夹路径，是指定的存放导出表格的文件夹。 */
data temp;
set &infile.;
keep &varlist. &var.;
if missing(&var.) then output temp;
run;
/* 注释：读入输入数据 &infile 之后，输出变量 &var 缺失的行至数据集 temp，并保留 &var 和 &varlist 变量 */
proc sql noprint;
select distinct count(*) into: count from temp;
quit;
%if &count>0 %then %do;
proc export data = temp outfile = " &outfile. \
&checkid._&checklabel._%left(&count)_&sysdate..xlsx" dbms=excel label replace;
run;
%end;
proc delete data=temp.;run;
/* 注释：通过 temp 数据集计算变量 &var. 的缺失个数，如果个数>0，则生成命名为"数据核查编号_核查名称_缺失个数_系统日期" excel 表存储至 &outfile 的地址，并删除 temp 数据集以方便下次的运算。 */
%MEND;
```

下例为调用该宏程序核查表 1"出生日期 BRTHDAT"变量是否存在缺失。如果存在缺失，则导出 excel 表格，列出该条记录的患者唯一编码（PID）、项目点编号（SITEID）和问卷填写的日期时间（QDATETIME），供后续进一步核实。假设存在 10 条缺失记录，导出的表格将命名为"001_出生日期有缺失_10_程序运行日期"，保存于"D:\项目数据核查结果"文件夹下。

```
%CHECK_MISSING(infile=dataset,var=BRTHDAT,varlist=PID SITEID QDATETIME,checkid=001,check_label=出生日期有缺失,outfile= D:\项目数据核查结果);
```

(2)检查数值变量是否超过允许范围的宏程序

```
%MACRO CHECK_LIMIT_NUM(infile=,var=,lowlimit=,uplimit=,varlist=,checkid=, checklabel=,outfile=);
/* 备注:lowlimit 和 uplimit 为"允许的正常值下限"和"允许的正常值上限"分别用以制定允许录入的正常值上下限范围。*/
data temp;
set &infile.;
keep &varlist. &var.;
if not missing(&var.) and (&var.<&lowlimit. or &var.>&uplimit.);
run;
/* 注释:读入输入数据 &infile 之后,在变量 &var. 不缺失的观测中,输出小于 &lowlimit 和大于 &uplimit 的行至数据集 temp,并保留 &var 和 &varlist 变量*/
proc sql noprint;
select distinct count(*) into: count from temp_&checkid.;
quit;
%if &count>0 %then %do;
proc export data=temp
outfile="&outfile.\&checkid._&checklabel._%left(&count)_&sysdate..xlsx"
dbms=excel label replace;
run;
%end;
proc delete data=temp;run;
%MEND;
```

下例为调用该宏程序核查表 1 中"收缩压 SBP"变量是否超出项目规定的合理值范围(50mmHg ~ 280mmHg)。如果存在异常,则导出表格,列出该条记录的患者唯一编码、项目点编号、问卷填写的日期时间、舒张压(DBP)、年龄(AGE),供后续的进一步核实。

```
%CHECK_LIMIT_NUM(infile=dataset,var=SBP,lowlimit=50,uplimit=280,varlist=PID SITEID QDATETIME DBP AGE,checkid=002,checklabel=收缩压异常,outfile=D:\项目数据核查结果);
```

在实际工作中,绝大多数逻辑核查需求均可归类后通过宏程序进行实现。对于较为特殊的情况,例如复杂变量衍生计算的复核,可以单独撰写核查程序。

3. 发送核查邮件

当完成数据核查任务时,可利用如下程序实现自动发送带有简要报告的电子邮件。由于 SAS 发送邮件需要配置 SMTP 服务器和发件人的地址,数据管理员可以在电子邮件客户端查找。邮件中需要列出的一些统计数据,可在 SAS 中计算得到。

```
OPTIONS Emailauthprotocol=login
Emailsys=smtp
Emailport=25
Emailhost="smtp.qiye.163.com" /* 调用的发件邮箱的 SMTP 地址*/
Emailid="example@163.com" /*调用的发件邮箱地址*/
Emailpw="PasswordEaxmple";/*调用的发件邮箱登录密码*/
filename mymail email" example@163.com "
subject="SAS OUTPUT SYSTEM"
encoding=gb2312;
Data _null_;
file mymail
to=("receiver@163.com") /*收件邮箱地址*/
subject="项目数据核查进展 &sysdate."
put "您好:";
put "【 left(&sysdate.) xxx 项目数据核查已完成】";
run;
```

4. 任务自动化

为解决手动运行 SAS 程序的繁琐,可以利用 Windows 操作系统自带的"管理工具"-"任务计划程序"功能,实现 SAS 软件的自动运行。通过设置"触发器",可以自定义运行的次数、频率、生效日期。另一核心设置是在"操作"模块新建指定的 .bat 脚本。该脚本可包含多条信息,将依次运行指定的数据核查程序。脚本内容见下例:

```
"C:\Program Files\SASHome\SASFoundation\9.4\sas.exe" -SYSIN %启动 SAS 软件%
"D:\项目数据核查程序\editcheck.sas" -NOSPLASH -LOG %SAS 逻辑核查代码目录%
"D:\项目数据核查结果\log.log" -PRINT
"D:\项目数据核查结果\log.lst" %打印运行日志并保存%
```

讨 论

数据质量是一项医学研究能否成功的重要依托[2]。对数据进行逻辑核查,可以帮助研究中心及时发现错误,即时解决疑问数据,使得疑问数据解决从研究接近结束时提前到研究进行中,加快研究进程,提高数据质量[3]。EDC 系统的兴起和应用,可以协助研究者较为便捷地实现这一工作。然而一些规模、经费和

(下转第 144 页)