数据采集与管理：数据库和数据文件

第一节 数据库基本概念

数据库是存放数据的地方。比起 Excel，数据库的优势是可以存放大量的数据并允许很多人同时使用里面的数据。有了数据库后，我们可以直接查找数据。

一、关系数据库

数据库有很多种类，最广泛使用的是关系数据库。关系数据库是由多个表组成的。Excel 是一张一张的二维表， 每个表都是由行和列组成的。同样，关系数据库里存放的也是一张一张的表，各个表之间有联系。简单来说：关系数据库等于多张表加各表之间的关系。

建立数据库，首先是构建每一张表的结构，每个表由一个名字标识。表包含带有列名的列，和记录数据的行。如下面这个表 4-1 名称为 GENERAL，里面有 10 列， FTYPE、FMID、SUBJ、NID、…… 为列名，又称为字段。每一行记录着每一位研究对象的信息，称为记录。各记录之间字段取值不一样，有变化，因此统计上又把字段称为变量。

表 4-1 GENERAL

FTYPE	FMID	SUBJ	NID	SEX	AGE	HT	WT	SBP	DBP
0	1	1	11	1	17.3	1.64	58	123	62
0	1	3	1	1	44.8	1.7	69.5	129	60
0	1	2	2	2	43.1	1.5	47.8	99	70
0	2	5	1	1	38.9	1.65	53	117	60
0	2	4	2	2	35.9	1.59	52	105	56
0	3	6	2	2	36	1.57	46	126	66
0	4	9	13	1	39.5	1.64	62	123	56
0	4	10	14	1	30.1	1.68	59	124	69

再看另外一个表 4-2，名称为 GENOTYPE, 字段有 SUBJ、A11、A12、… 等。

表 4-2 GENOTYPE

SUBJ	A11	A12	A21	A22	A31	A32
1	20	10	B	A	G	R
3	10	20	A	B	R	G
2	10	10	B	A	R	R
5	10	10	B	B	G	R
4	10	10	B	B	R	R

6	10	20	B	B	R	G
9	10	10			R	R
10	10	10	B	B	R	G

这两个表通过 SUBJ 关联起来，给定一个 SUBJ，在 GENERAL 表里可以查到他的基本信息和表型信息，在 GENOTYPE 里可以查到他的基因型信息，在数据分析时，如要把基因型信息与表型信息关联起来分析，就需要通过 SUBJ 将每一个需要的变量从各表里提取并串起来，形成一个新的供统计分析用的表，将它保存为一个单独的文件，称为工作数据文件。

二、数据采集和管理系统

用来管理数据库的计算机软件（关系数据库管理系统）有很多种，比如 MySQL、Oracle 等。SQL 是为操作数据库而开发的一种语言，它可以对数据库里的表进行操作，如存储、修改，查找等。

数据采集和管理系统顾名思义是为采集数据和管理数据设计的一套系统，它包含数据库数据表、前台数据采集页面和用户操作页面、后台程序。后台程序是用来处理用户指令，调用 SQL 语言与数据库沟通，以实现保存、更新、查找数据以及对数据进行统计分析和报告等。数据库里的数据表设计需要有经验的研究人员参与，详见“工作数据文件设计原则”。用户操作页面的设计更需要有现场工作经验的研究人员参与，以适应现场工作场景。仅仅靠计算机编程人员很难做出有实用价值的数据采集和管理操作系统。

第二节 数据库设计

数据库设计包括组成数据库的表、各表包含的字段、表之间通过什么字段相联系。

一、课题要收集的变量

在设计一个研究项目的数据库时，首先要确定要收集哪些变量。任何一项研究需要收集的变量无外乎三大类：

①一般资料：包括研究本身相关的变量（如研究中心编号、地点、研究对象分组等）与研究对象的基本特征（如性别、年龄、职业等）；

②与要研究的危险因素有关的变量；

③与要研究的结局有关的变量。

一项大型研究有主要目的和次要目的，有主要观察结局变量和附加观察变量。通过查文献、结合专业知识与实践经验首先确定要收集哪些变量和如何收集这些变量。有些变量是通过问卷询问得到，有些变量

是通过体格检查得到，有些是通过实验室检测生物样本得到。如通过体检得到，就要明确检查所用的仪器或工具，如测血压所用的血压计类型与操作方法，特殊的仪器要明确型号和生产厂家。如通过生物样本检测获得数据，要明确采样时间和方法、生物样本储存和运输过程、以及实验室具体检测方法等。

二、表单设计原则

根据数据采集的流程和现场场景设计数据库表，如一项问卷为一张表，同时完成的体检项目（如身高、体重、血压）为一张表，一项实验室检测获得的系列数据（如生化检测）为一张表。要求同一位研究对象的一个表的数据，可由一个人一次完成资料的采集与录入。

设计数据表时要遵循简单性原则，即只记录原始变量，体现在两个方面：

1. 避免重复。如通过二次计算生成的变量不列入数据库中。可以想象如果一个数据库中既有身高（HT）、体重（WT）两变量，又有由此两者计算出来的体重指数（BMI）变量，当人们拿到这样一个数据库时，就会问：这个 BMI 是什么？能不能直接用（要不要检查一下这个 BMI 计算的有没有错误）？如果发现有 BMI 与 WT 和 HT 计算不符怎么办？可想而知，一个不必要的重复既导致很多无意义的工作，又带来混淆和不知所措。

2. 避免叠加。如变量“服药至出现皮疹时间”，在实际操作时出现很多问题：有的患者没服药，怎么记录？有的患者服药了但没有皮疹，怎么记录？有患者没有服药却出现了皮疹，怎么记录？解决办法是记录最原始的信息：可以把这个存在叠加信息的变量拆分为：是否服药（0=否，1=是），是否出现皮疹（0=否，1=是），服药至出现皮疹时间（连续变量）；若病例资料中没有现成的“服药至出现皮疹时间”，则需要再拆分成两个变量：开始服药日期（时间变量），出现皮疹日期（时间变量）。

从上述内容可以发现，确定数据库收集哪些变量，要求研究设计者对于课题研究目的和内容、研究方案和实施过程各环节都有深入的了解和体验，并在实际实施过程中不断完善。

第三节 数据文件行列设计

一个数据文件相当于数据库中的一个表，基本结构为行×列表的矩阵形式。行列设计有两种基本格式，一是结构化数据，二是非结构化数据。

一、结构化数据

结构化数据每一行为一条观察记录，每一列代表一个观测变量。可直接使用变量名对观测指标进行统计分析。对于随访性研究，每个研究对象有多次观察，其数据文件的行列设计有两种基本类型：

1、横向数据

一个研究对象一条记录，同一观察指标不同时间的观测值用不同的变量名记录在同一行不同列中，如下表 4-3 所示。适用于在固定时间点进行重复测量的随访研究，即每个观察对象的随访次数与每次随访的相对时间是固定的。

表 4-3 体检结果表（横向数据）

ID	Sex	Date1	Lgth1	Lbs1	Date2	Lgth2	Lbs2	Date3	Lgth3	Lbs3
2540	1	12170	23	10.4	12194	24	11.6	12255	26	14.4
2630	2	12165	23	9.8	12205	25	13.3	12267	26	17.3
2740	1	12165	22	8.7	12197	23	11.8	12261	26	15.0
2840	1	12164	21	8.0	12198	22	10.8	12260	25	14.6
3040	1	12187	22	10.9	12206	23	11.7	12380	28	18.1

上表中 ID 为研究对象编码，Date1、Date2、Date3 分别表示第 1、2、3 次测量时间，Lgth1、Lgth2、Lgth3 分别表示第 1、2、3 次测得的身长，Lbs1、Lbs2、Lbs3 表示测得的体重。一般来说，横向数据仅出现于数据分析用的工作数据。在数据库表设计中通常不会用横向数据，试想要等到一个研究对象完成全部随访后，才能完成一条记录的数据的填写与输入，既不利于现场操作又容易出错。

2、纵向数据

一个研究对象每一次观测为一条记录，每个研究对象有几次重复观测就有几条记录。同一个观测指标用相同的变量名记录在不同的记录中。如上表中，研究对象编号为 2540、2630 的三次随访记录结果如下（表 4-4）。

表 4-4 体检结果表（纵向数据）

ID	Sex	Date	Lgth	Lbs
2540	1	12170	23	10.4
2540	1	12194	24	11.6
2540	1	12255	26	14.4
2630	2	12165	23	9.8
2630	2	12205	25	13.3
2630	2	12267	26	17.3

纵向数据既适用于固定时点的随访，又适用于非固定时点的随访。数据库中的随访记录表用纵向结构，每完成一次随访即时完成一条记录的录入。用于统计分析的工作数据文件也需要纵向数据，通常用于重复测量数据的回归模型，如广义估计方程（GEE）和混合模型（Mixed），都要求输入纵向数据。纵向数据工作文件包含重复测量的指标（如上表的 Date、Lgth、Lbs）和研究对象特异的指标（如上表的 ID、SEX）。前者每条记录不同，后者每个研究对象不同。

二、非结构化数据

结构化数据每一列代表一个观测指标，非结构化数据则不然，不同观测指标的观测值用同一个变量记录在同一列但不同的记录（行）中，如下表的 Value 变量，另有一个变量代表所观测的指标名称，如下表的 Test 变量。上例中研究对象 2540 的三次两个指标的随访记录列表如下表 4-5。

表 4-5 体检结果表（非结构化数据）

ID	Sex	Date	Test	Value
2540	1	12170	Lgth	23
2540	1	12170	Lbs	10.4
2540	1	12194	Lgth	24
2540	1	12194	Lbs	11.6
2540	1	12255	Lgth	26
2540	1	12255	Lbs	14.4

当随访时间和检测指标都不固定时，适用非结构化数据存储数据。临床电子病例中的实验室检测资料，多为非结构化数据。非结构化数据需要重新整理、提取、生成结构化数据，才能适用于统计分析。易侕统计软件“非同步重复测量数据转换”模块即用于将非结构化数据转换为结构化数据。易侕大数据整理系统也有专用的“非结构化数据提取”模块。

第四节 工作数据文件

研究对象的信息按性质和类别，存储在数据库不同的表里，通过一个编码可以将同一个研究对象的用于统计分析需要的信息，从不同的表里提取并串联起来，形成一个工作表，存成一个工作数据文件。工作表里的列名称就是分析用的变量名，每一行为一条记录。保存文件格式为文本文件，字段（列或变量）之间用制表符、逗号或空格分隔，常用制表符分隔。选择分隔符的时候要注意字段取值里是否可能包含分隔符，特别是当字段取值有文字描述时，里面有可能包含逗号、空格或制表符。如果有，读取数据时字段取值就会被分隔成两个或多个字段，导致字段错位，数据读出来杂乱无章，无法进行统计分析。

一、工作数据文件设计原则

用于统计分析的工作数据文件设计应遵循以下原则：

- 非随访数据：每个患者一行，每个变量一列；
- 随访数据：对患者在不同时点进行测量，每个患者一个唯一的编号（ID）；每次测量一行；变量中必须有每次测量的时间（具体日期或者时点）；
- 连续变量（如年龄）用原始数值记录，不加单位；
- 分类变量（如血型、分组）用 0,1,2 表示，不用中文和字母；

- 分组变量对照组编码为“0”；
- 变量名尽量简短，变量名最好仅由英文字母和数字组成，需用英文字母打头，不能有空格或运算符号如 +、-、*、/、%、^、&等；
- 缺失变量用空格或者“NA”表示；
- 不参与分析的变量（如单位名称）剔除；
- 参与分析的变量都要数字化（赋值中不能含中文或非数字字符）；
- 随访结局变量通常编码为：0=未发生结局，1=发生结局。

例，下面这个 Excel 数据表 4-6, 目的是为了研究肝癌患者化疗后皮疹情况和生存的关系，表结构和前三条记录如下：

表 4-6 原始资料表（不能直接用于数据分析）

姓名	性别	年龄	状态	死亡时间	死亡原因	病因	分期	病灶数量	负荷	肿瘤分布	血管侵犯	皮疹	
												服药至出现时间	最重级别
王某	男	42	死亡	2011/9/29	消化道出血	乙肝	A	1	≤50%	单叶	无	无	0
刘某	女	46	死亡	2011-3-13	呼衰	乙＋丙	B	1	>50%	单叶	有	46	1
全某	男	43	存活	2010/11/17	其它死因	其它	A	2	无法判断	双叶	无	无	0

上表不能直接用于数据分析，首先它不是一个行*列的结构，第一行列名称不能有复合列，列名称也即变量名需由英文字母和数字组成，不能含中文。其次用于分析研究对象的数据需要编码成数字，不能含中文。数据需做如下清理：

“姓名”是字符，没有分析的意义，可以删除或用患者编号代替；

“性别”通常编码为 1=男，2=女；

“状态”通常编码 1=死亡，0=存活；

“死亡时间”是时间变量，格式需要统一到：YYYY-MM-DD；

“死亡原因”可以有两种方式：①用一个变量“死亡原因”表示（1=消化道出血，2=呼衰，3=其它死因）；②用两个变量分别表示：“是否死于消化道出血？”与“是否死于呼衰”，编码为：0=否，1=是；

“病因”可以有两种方式：①用一个变量“病因”表示（1=乙肝，2=乙+丙，3=其它）；②用三个变量分别表示：“乙肝”、“丙肝”、“其它”，编码为：0=无，1=有；

“分期”可以编码为：0=A，1=B；

“负荷”可以编码为：0=小于等于 50%，1=大于 50%，NA=无法判断；

“肿瘤分布”可以编码为：0=单叶，1=双叶；

“血管侵犯”可以编码为：0=无，1=有；

皮疹“服药至出现时间”需要用两个变量“出现皮疹”（0=无，1=有；如需要考虑皮疹严重程度可以编码（0=无，1=轻，2=中，3=重）和“服药至出现皮疹时间”（连续变量，如未出现皮疹者该变量缺失）；

皮疹“最重级别”，未出现皮疹者该变量为空格或 NA。

变量名只占一行，不分两行。

整理好的《数据》如下表 4-7：

表 4-7 数据表（可用于数据分析）

ID	SEX	AGE	DETH	DETHD	REASON	DIS	BCLC	NS	FUHE	FENBU	INVES	RAS	RAST
1	1	42	1	2011-9-29	1	1	0	1	0	0	0	0	NA
2	2	46	1	2011-3-13	2	2	1	1	1	0	1	1	46
3	1	43	0	2010-11-17	3	3	0	2	NA	1	0	0	NA

数据整理成上表后，单纯从变量名上看不知道其代表什么观测指标，从赋值上不知道其代表什么意义，这就需要有另一个“变量说明”文件来记录变量名与赋值编码的含义。

二、变量说明文件制作原则

独立于原始数据文件新建一个 Excel 文件进行变量说明，最后保存为制表符分隔的文本文件，如上表中的前 4 个变量的说明如下：

表 4-8 变量说明表

变量名	取值编码	意义
ID		Subject identification
SEX		Gender
	1	Male
	2	Female
AGE		Age, years
DETH		Death or survival
	0	Survival
	1	Death

“变量说明”要点如下：

- 包括“变量名”、“取值编码”和“意义”三列；

- “变量名”同数据文件的列名称；
- “取值编码” 与数据文件对应的列数据编码一致， 分类型变量列出编码，连续型变量此处空白。分类变量编码从变量名对应的第二行开始顺序填写，一般用 0,1,2……顺次编码，对照组常规编码为 0；
- “意义”表示变量名和取值编码代表的意义，如上例中 SEX，表示研究对象的性别（gender）,编码 1 表示男（Male），2 表示女（Female）；若文章输出图表需要显示为英文，则“意义”为英文；若文章输出图表需要显示为中文，则“意义”为中文。

易侕统计软件可以直接读入“变量说明”文件，也可手动对每个变量进行注解，在输出图表中自动将变量名替换成注解。

如用易侕 DataWeb 构建数据库录入数据，最后可根据需要，从各表单中提取所要的变量，导出分析用的数据文件，并同时自动输出变量说明文件。

易侕 DataWeb 数据系统

易侕 DataWeb 是集近百项大型流行病学研究的数据采集和数据管理经验开发的，是一个网页版的科研项目数据录入和管理系统，可在电脑、平板、手机上运行。其功能设置既贴切现场工作的需要又能满足课题负责人、数据管理员和数据分析人员的各种操作需求。易侕 DataWeb 有如下功能特点：

1. 具备自主创建功能。易侕 DataWeb 是一个开放（不是开源）的系统。用户可以添加自己的研究项目、表单和问项，设定参与研究的工作人员，并对人员设置权限。

2. 具备自主修改表单功能。在实施过程中可以更新课题方案、表单和问项，既保证课题方案的严格规范，又保持其灵活性和实用性。

3. 用户可以选择将自己设计的表单和问项共享给他人，有利于指标的统一和问卷的标准化，适合课题之间资源共享和数据合并。

4. 权限管理，可自主设置权限，区分录、读、改权限。通过添加工作人员和对其权限设置可以实现多科室、多中心大项目数据采集和管理。

5. 简单实用的数据录入自动检错、数据查询、数据质疑和更正功能，提高现场工作效率。

6. 质控是关键，详尽的进度报告，随时查看数据质量。即刻生成进度报告，内容包括各变量的数据分布、以表单为单位的进度统计和以研究对象为单位的资料完整情况统计，即时跟踪研究进度和数据质量。

7. 导出结构化数据，无需再花时间整理数据。既有一键下载全部数据的功能，又有跨各表单的数据提取和合并功能，可即时生成用于统计分析的工作数据文件。

8. 简便灵活，大小项目均支持。支持多终端录入，手机、iPad 或电脑录入与管理数据，功能强，极简操作，不需要专门培训。可以用于单中心或多中心复杂项目。

使用易侑 DataWeb 首先要注册获得一个用户编号。官网为 http://www.empowerstats.net/dataweb/。一个用户既可以是自己登记的课题的负责人，又可以是其他人登记的课题的工作人员。用户登陆后，可看到自己负责的课题和参与他人的课题。

用户有两条主线进入易侑 DataWeb 功能页面，一是通过用户名下拉菜单【我的课题】、【我的表单】、【我的问项】进入到课题、表单、问项的查阅和设计页面；二是在“课题列表”里选择课题，进入课题数据应用页面，包括数据录入或导入、查阅、提取、质疑、进度报告和数据下载等。

第一节 项目数据库设置

一、表单

（一）基本概念

课题由若干个表单组成，表单由问项（问题或检测指标）组成。表单是根据数据采集方法与过程来设计的。通常将同一过程或方法收集到的数据存放到一个表单中，以方便填写。一个表单要求一次性完成资料采集与同步录入，如一次完成的问卷可以做一个表单，一次完成的随访记录做一个表单。表单可以重复使用，如一个项目对研究对象进行多次随访，每次随访询问内容和检查项目相同，就可以用同一随访表单。不同的项目也可以用到同一表单。

每个用户都可以在易侑 DataWeb 里创建自己的表单，通过【创建表单】菜单创建新表单，通过用户名图标下的【我的表单】查找并修改自己创建的表单。用户可以把自己的表单共享出来，一旦系统检测到用户的表单共享出来后被人引用，系统自动关闭对该表单的修改按键，用户不能再对之进行修改，否则就会影响他人的数据采集。

（二）创建表单和呈现表单页面

下图 5-1 为创建新表单示例：

上图中的“表单设计”框不是手动填写的，而是通过点击“表单设计”按钮，进入表单设计页面，在表单设计页面完成设计并保存后，系统自动填写表单设计框。下图 5-2 是表单设计页面示例：

表单是由问项组成，输入问项的方法有：（1）如知道问项编号，在问题编号处填写问项编号，直接导入问项；（2）输入关键词，点击左边的搜索按钮，查找问项库，然后调入所要的问项；（3）点击左边的编辑按钮，调入添加问项页面，详见下面的“添加问项”部分。

每个表单的呈现有 5 种页面格式，可根据自己的喜好切换。不管用什么格式，录入到数据库里的数据都是一样的。下图 5-3 是表单示例：

二、问项

（一）基本概念

数据库是由一个个表单组成，而表单又是由问项组成的。问项指问卷所问的问题或检测项目所检测的指标。问项最终形成数据库中的字段或称变量。问项重复使用很常见，如基本上每个项目都要收集性别，“性别”这个问项就会重复用在多个课题中。

每个用户都可以在易佾 DataWeb 里添加自己的问项，通过【添加问项】菜单添加新问项，通过用户名图标下的【我的问项】查找并修改自己添加的问项。用户可以把自己的问项共享出来，一旦系统检测到用户的问项共享出来后被人引用，系统自动关闭对该问项的修改按键，用户不能再对之进行修改，否则就会影响他人的数据采集。

（二）添加问项

下图 5-4 是问项的设计页面示例：

问项设计的关键字段是问题描述和题型，题型分：数字、单选、多选、文字、日期、时间、其它。如为数字类型，页面自动带出最小值、最大值、数值单位三个填空项，给定最大值和最小值后，在数据录入时系统会自动根据最大值和最小值检错。如为日期型，数据录入窗口会自动添加日历按钮方便用户录入。如为单选或多选题，在问题描述框要输入选项，先输入问题描述，然后另起一行，输入选项，每个选项另起一行，系统通过换行符自动识别问题描述与各选项。如下图 5-5 所示：

多选题区别于单选题，单选题在数据库中用一个变量表示，而多选题在数据中转换成多个两分类变量，每个变量编码为 0 表示无 （未选），1 表示有 （选）。易佾 DataWeb 在数据录入时，自动将多选题选择结果转换成多个 0/1 两分类变量。

特别要注意的是，在设计多选题时，第一个选项要为“全无”或“全否”，这样只要回答了此问题，就至少有一个选项被选择。如果选了“全无”或“全否”，则其它选项自动为无或否。如果所有选项均未选，则表示研究对象没有回答此问题，数据为缺失。有了“全无”选项才可以区别”未答“与”所有选项均为否”。

（三）变量赋值

各题型对应的变量在 DataWeb 系统内部赋值为：（1）填空题：包括数字、文字、日期、时间和其它题型，均为所填写值；（2）单选题：按选项顺序从 1 开始递增，依次为：1、2、……；（3）多选题：每个选项为一个变量，每个变量赋值为： 0=未选(No)，1=被选(Yes)。

数据导出值：即查阅数据列表时和下载的数据看到的值，填空题和多选题的导出值同内部赋值；为方便易佾数据分析，单选题导出值自动做了如下重新编码：（1）如为两分类选项：原编码 1 新编码仍为 1，原编码 2 新编码为 0；（2）如为多分类选项：原编码 1、2、3、……，对应新编码依次为：0、1、2、……。

三、课题

（一）登记课题

课题即研究项目，有其特定的研究目的和研究设计。研究设计体现在易佾 DataWeb 系统里就是课题的实施方案。易佾 DataWeb 是一个开放系统，每个用户都可以在易佾 DataWeb 系统里登记自己的课题，登记人即为课题负责人。下图 5-6 为一登记课题示例：

上图中的“实施方案”框不是手动填写的，而是通过点击“方案设计”按钮进入设计页面，完成设计并保存后，系统自动填写实施方案框。

（二）课题实施方案

实施方案包括课题所用的表单和随访时间表（如有随访），并明确关联各表单的研究对象编号和随访次序变量（如有随访）。下图 5-7 是方案设计页面：

通过查找表单库调入表单。数据是按表单存放的，数据分析时要将不同表单的数据按研究对象编号联系起来。系统会自动检测各表单共同的字段名，如上图中的“研究对象编号”，选择该字段表示系统可根据该字段识别来自不同表单的同一研究对象的数据。

同理，如果实施过程涉及多次随访，每次随访用到多个表单，要将每次随访的多表单数据串到一起，不仅需要“研究对象编号”变量，还需要“随访次序或时间”变量。

（三）课题工作人员和权限

课题工作人员包括参与资料采集、录入、查阅、修改、提取和随机化控制员（如为 RCT 研究）。课题负责人可以为本课题添加工作人员，并指定每个工作人员的权限。工作人员权限包括（1）对课题所用的各表单录入、查阅和修改；（2）添加人员；（3）生成随机数并对入选者随机分组。

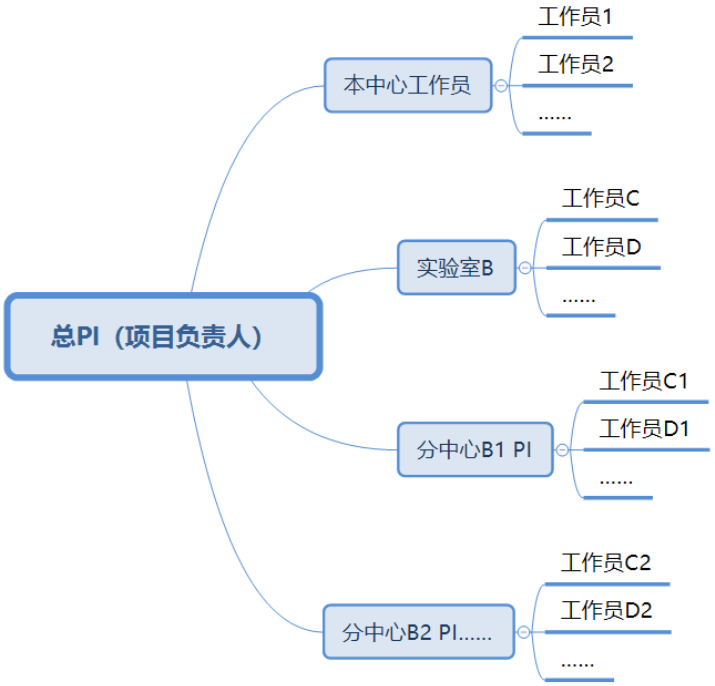
易侑 DataWeb 不仅适用于单中心研究，也适用于多中心研究项目的管理。如一个工作人员有添加人员的权限，则可以对其所添加的人分配自己所拥有的权限。通过权限设置，实现各科室数据的条条管理（对特定表单的权限）和分中心数据的块块管理（对各表单的本中心录入的数据的操作权限）。多中心项目的操作流程如下：

第一步：由课题负责人（PI）登记课题。

第二步：由 PI 登记本中心工作人员和分中心 PI。

第三步：由 PI 为本中心工作人员授权，为分中心 PI 授权。本中心工作人员权限仅是对各表单的录、读、改权限。分中心 PI 的权限不仅有对各表单的录、读、改的权限，还包括添加工作人员的权限。

第四步：当分中心 PI 被授予了添加工作人员权限后，就可以添加分中心的工作人员，并设置分中心每个工作人员对各表单的录、读、改权限。



1. 添加人员

课题负责人在添加人员时，首先每个工作人员注册易侕 DataWeb（免费注册），每工作人员首先要获得一个用户编号后，然后才能添加到课题中。添加人员操作如下图 5-9 所示，填写工作人员编号后，系统能自动调出其姓名、邮箱等信息。

2. 设置权限：

权限包括对各表单的录、读、改权限、添加工作人员的权限、随机控制（如为随机化对照的临床试验研究）权限。如被授予添加工作人员的权限，即被指定为分中心负责人或某部门/科室的负责人，分中心负责人负责添加本中心工作人员，分部门/科室的负责人负责添加本科室的工作人员。随机化对照的临床试验的随机控制员的权限是负责生成随机数，以及按提交的研究对象编号次序分配随机数。权限设置的基本原则是：

i. 分中心 PI 对其添加的工作人员授权时，能授予的权限仅在本人所拥有的权限内。

例：总 PI（**A**）只授予了某实验室负责人即分中心 PI（**B**） 对实验室数据表单 F1 的录、读权限，B 添加了 C、D 两个工作人员，B 最多只能对 C、D 授予对表单 F1 的录、读权限。因为 B 没有对表单 F1 的修改权限，所以由 B 添加的 C、D 两人也不可能有对表单 F1 的修改权限。

ii. 分中心 PI 的权限自动限制在对由本人及由其添加的分中心工作人员录入的数据的操作。

例：总 PI（**A**）添加了 **B1**、B2 两个分中心 PI，**B1** 分中心添加了 **C1**、**D1** 两个工作人员，B2 分中心添加了 C2、D2 两个工作人员。虽然 B1 有对数据表单 F1 的录、读、改权限，但只能读、改 B1、C1、D1 录入的数据。看不到其它分中心录入的数据。

工作人员对各表单的权限是随课题而不同的。如两个不同的课题都用到同一表单 F1，某用户 A 参与了这两个课题，A 在一个课题里对 F1 权限与在另一个课题里对 F1 的权限可有不同（图 5-10）。

四、中央随机

一个实用和有效的中央随机系统是严格执行随机分组的保障。在了解易侑 DataWeb 中央随机系统之前，首先有必要理解随机分组过程中，可能出现的偏性和如何控制这种偏性。

随机分组过程中偏性控制的基本原则是入选时不知道分组，即先入选后分组。这样研究人员在掌握入选标准的时候才不可能有偏性。

（一）单纯随机(simple randomization)

操作流程：来了一个病人，如符合入选标准，入选并分配研究对象编号，然后随机分组。这是最基本也是最常用的方法。

单纯随机的缺陷是不能保证每组样本量相同。单纯随机分到各组的人数可能相差比较大，当样本量比较小的时候，会大大降低检验效能。如样本量是 10，分到一组 7 个另一组 3 个的可能性还是比较大的。

单纯随机理论上入选时不知道分组，但实际操作时，人们可以根据以前的分组情况判断下一个病人分组概率。假如实验分 A、B 两组，前面连续入选了 4 个人都分到了 A 组。这时候你就可以猜测下一个病人分到 B 组的可能性很大。因为随机分组的结果应该是出现 A 或 B 的概率一样。然而，这种预知下一个很可能会分到哪一组的情况是少数，而且也不能保证绝对正确，不会对结果带来很大的偏性。

（二）区组随机(block randomization)

区组指的是由几个研究对象组成的一个整体。区组大小指的是区组内的研究对象人数。区组大小取试验组数的倍数。如试验分 A、B 两组，区组大小可以是 2、4、6、8……。如试验分 A、B、C 三组，区组大小就应该是 3、6、9……。区组随机是将区组内的病人随机并且均匀地分到各试验组中，分到每个试验组的人数相同。如试验分 A、B 两组，区组内有 4 人，按入选顺序有 AABB、ABAB、ABBA、BBAA、BABA、BAAB 六种分法，取哪种分法由随机数决定。

操作流程是：来了一个病人，如符合标准就入选，然后入选第 2 个……，等到一个区组的所有病人都入选齐了后，再随机分组。

当区组(block)比较小的时候，区组随机的规律可以被利用。如试验分 A、B 两组，两人一个区组，只有 AB、BA 两种分法，如果前面几个区组随机结果都是 AB，那么下一个区组随机的结果是 BA 的可能性很大。与单纯随机一样的，它的规律有可能被人利用。如果区组比较大，如四人一个区组，有六种可能分法，能猜测到下一个区组是哪种分法的可能性小很多。

区组随机的缺陷是要等一个区组的病人全部入选以后再随机分组。否则根据区组内前面的病人分组情况就能确定后面病人的分组。如试验分 A、B 两组，4 人一区组，前面两人都分到 A 组，那么后面两人一定都是 B 组。这里的关键是能不能等，根据试验性质，有时候不能等。如罕见病，可能很久都等不到一个病人。比较紧急的病，入选后必须马上进入治疗/处理，也不能等。有些病当区组人数比较少的时候，等待时间不会很长，可以等；但如区组比较大就不能等。而当区组比较小的时候区组随机的规律比较容易被利用，也不能保证入选时不知道分组。

（三）随机区组的区组随机(block randomization with random block size)

有没有办法能同时满足三点：（1）入选的时候不知道分组；（2）尽可能随机分到各组的人数相同；（3）入选一个分组一个，不需要象区组随机那样等区组全部完成入选后再分组？答案就是：随机区组的区组随机。这里的随机区组指的是区组大小是随机的。

操作流程如下：

第一步：产生随机区组大小系列，即区组的大小是随机数组成的系列。假设某试验分 A、B 两组，生成随机区组大小系列如：2、4、2、6、4……。当然每个随机数都是试验组数的倍数。

第二步：按区组随机的方法，对每个区组进行随机分组。如上例随机分组的结果是：第 1 区组（2 人）随机结果为 AB；第 2 区组（4 人）随机结果为 AABB；第 3 区组（2 人）为 BA，即第 4 区组（6 人）为 ABBABA，……。

第三步：将第一步的随机区组大小系列销毁，将第二步生成的随机分组系列连起来，密封保存，以后按该序列依次对每一个入选病人分组。

随机区组大小的区组随机不能保证每组人数完全相等，因为试验中止的时候，最后一个病人所在的区组不一定全入选完，然而最大的组间人数差异只取决于用到的最后一个区组的大小。

随机区组大小的区组随机实施要点是：不能保存第一步的随机区组大小系列，否则实施过程中就可根据区组大小确定下一个病人会分到哪一组。对第二步生成随机分组系列要密封保存，否则知道随机系列就知道下一个病人会分到哪一组。这两点用中央随机系统是很容易实现的。

（四）易侑 DataWeb 中央随机系统设计要点

1. 一个项目有一个指定的随机控制员，不能是 PI 本人，不能录、读、改表单数据，确保随机控制员是独立的和中性的。

2. 随机控制员预先根据课题设计，产生随机系列，自动封存于系统内。产生随机系列的方法默认为随机区组大小的区组随机，也可以指定固定区组大小的区组随机。方法不同，以后实行过程中对病人进行入选与分组的操作不同。

3. 如采用随机区组大小的区组随机，每入选一个病人就可以即时将该病人分组，由随机控制员输入病人编号，才能调出分组代号。

4. 如采用固定区组大小的区组随机，每完成一个区组全部病人的入选才能对该区组所有病人分组，由随机控制员输入一个区组所有病人的编号，才能调出该区组每个病人的分组代号。

5. 用分组代号（如颜色）代替试验组名称（盲法），仅随机控制员才能看到分组代号与试验组的对应关系。

6. 自动保留完整的病人编号与对应的分组记录，一个编号一旦被分组，不能被更改和删除。

由上可知，使用易侑 DataWeb 中央随机系统可以保证研究过程中的入选与分组严格按流程操作，随机系列严格保密，操作过程有完整记录可追踪。

第二节　项目数据库应用

一、数据录入和数据导入

数据录入指按表单逐条记录录入数据。表单的第一项“研究对象编号”录入框有一搜索键，录入员在输入研究对象编号后，可以点击搜索，看该编号是否已有记录，以防止重复录入。数据录入页面示例如下图 5-11：

数据导入是将一批数据一次性导入到易侑 DataWeb 的某一表中。从外部文件导入数据，外部文件格式可以是：Excel 表单（只读单个表单）、制表符或逗号或空格分隔的文本文件(.xls, .csv, .txt)；或 EpiData (.rec) 数据文件。如为 .xlsx, .xls, .csv, .txt 文件，文件的第一行为字段（变量）名，第二行开始为数据记录。

给定外部数据文件后，系统会自动列出文件中的变量名与变量注解（如为 .rec 文件）和变量取值分布，用户需要逐一将外部文件的变量与编码与所选表单中的问项进行匹配，表单中如有问项在导入数据文件中找不到对应的变量，导入的记录该问项将被置为缺失。变量匹配页面示例如下图 5-12：

上图显示从 QGN.rec 文件导入数据，QGN.rec 中的变量 ID 对应于 Q1，A1 与问项 Q61 对应，A1 取值编码 1 对应于 Q61 的 2“经常”，2 对应于 Q61 的 3“有时”，3 对应于 Q61 的 5“从不”。如待导入的问项为连续性数字变量，在问项处会列出变量的取值单位、最小值与最大值，在匹配导入数据变量时，请检查变量单位与取值范围是否一致。匹配操作可以通过拖拽变量名与取值编码的方法完成。

完成表单问项与导入数据变量的匹配后，查看数据，待导入的数据将显示在下面的数据列表中。示例如下图 5-13：

或复制粘贴 (cntl+v) 数据 (请仔细核对字段名) 到下表:　查看数据　上传数据

	Q1	Q61	Q62	Q63	Q64	Q66	Q67	Q68	Q70	Q74	Q80	Q81	Q83	N
1	P00001	5		4	1	2	2	3	2	2	2	1	3	
2	P00002	3		1	1	3	2	3	3	1	5	4	1	
3	P00003	2		3	3	3	1	3	3	2	3	2	1	
4	P00004	2		4	1	2	2	3	4	2	4	2	1	
5	P00005	5		2	2	1	2	1	3	2	3	2	2	
6	P00006	5		3	3	3	1	2	4	3	3	1	1	
7	P00007	2		1	2	1	2	1	4	4	3	2	1	
8	P00008			4	1	2	2	3	2	2	3	1	3	
9	P00009	5		2	5	1	1	2	3	4	1	1	2	
10	P00010	2		5	2	1	2	2	1	3	3	3	1	

点击“上传数据”后数据才会正式保存到系统中。

也可将数据直接复制（ctrl+c)、粘贴（cntl+v）到上述的数据列表中，然后点击“上传数据”。需特别注意的是，分类型数据的取值编码在 DataWeb 系统内的内部赋值与下载的显示值不同。如问项 Q61 显示的数据取值编码为：0=总是，1=经常，2=有时，3=偶尔，4=从不；而在 DataWeb 数据库内部取值为：1=总是，2=经常，3=有时， 4=偶尔，5=从不。 两分类选题，在系统内部每个选项的取值为：1=是，2=否；数据导出后取值编码自动改为：0=否，1=是。待导入数据的值应为内部取值。因此数据导入时，通过变量与取值编码的匹配后，可保证导入数据正确赋值，如直接复制粘贴数据到列表中，则需要仔细核对变量取值是否与内部取值一致。

二、数据查阅与数据质疑

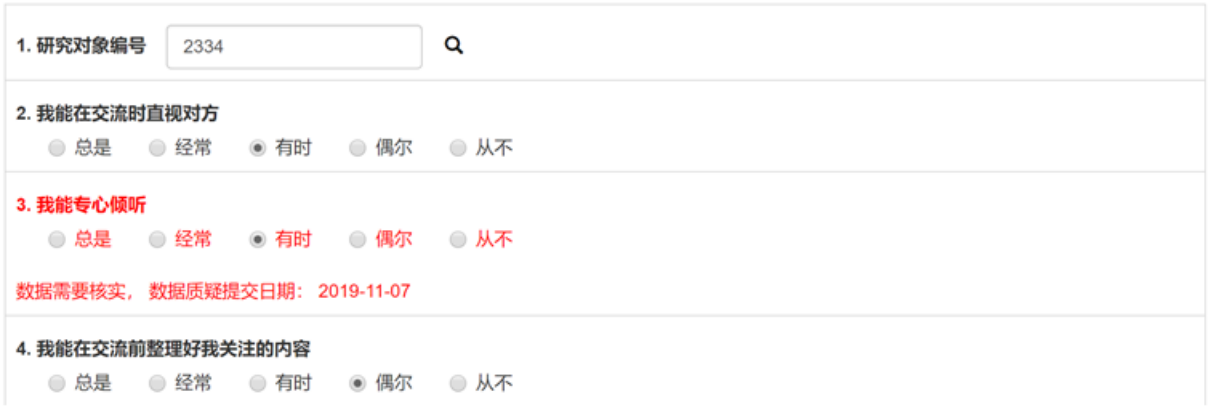
数据采集与数据分析过程中时常需要查阅数据，并对异常取值提出质疑要求核实。数据查阅示例如下图 5-14：

数据查阅: 素质/气质　营造良好氛围　保持对话题饶有兴致　读懂讲者　换位思考　专注与总结　理解并反馈　读懂讲者

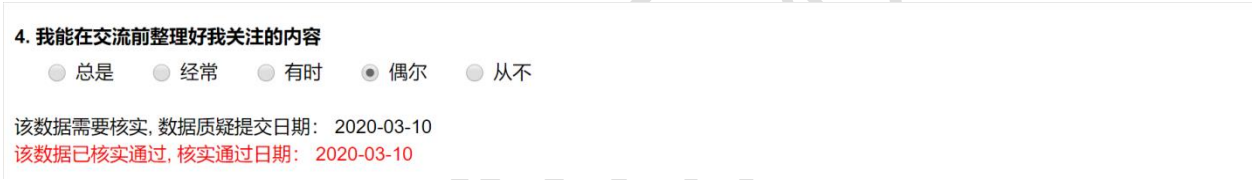
记录号	SID	REGDATE	REGUID	Q1	Q61	Q62	Q63	Q64	Q66	Q67	Q68	Q70	Q74	Q80	Q81	Q83
112470	2334	2019-11-07	1	2334	3	3	4	4	4						4	4

可通过设定研究对象编号、录入员的用户编号、录入日期缩小查阅范围。从查阅显示的列表中，点击记录号进入表单录入界面，如对某问项的取值提出质疑，需要数据采集/录入员核实，鼠标右击该问项，

在弹出的对话框提交质疑。被质疑需要核实的问项，显示为红色，并附有数据质疑提交日期。如下图 5-15 的问项 3 所示：



回复质疑操作亦是如此，查阅到该记录后，如被质疑的数据有输入错误，输入正确的数据，然后鼠标右击问项，在弹出的数据质疑对话框提交数据核实结果。提交后显示如下：



另一种途径提交数据质疑是提交数据质疑表。示例如下：

系统会自动根据研究对象编号查找数据录入员，如果是多中心研究，根据录入员编号核对是否是本中心的数据，只有对本中心的数据才能提出质疑。

回复质疑则可通过查看被质疑的数据，输入正确值来确认被质疑数据。示例如下图 5-18：

SID	表单	问项	取值	状态	提交日期	提交人	录入员	正确值
223356	素质/气质	我能在交流时直视对方: 0=总是|1=经常|2=有时|3=偶尔|4=从不	2	待核实	2020-03-11			
224	素质/气质	我能在交流时直视对方: 0=总是|1=经常|2=有时|3=偶尔|4=从不	4	已核实	2019-11-07			
2334	素质/气质	我能专心倾听: 0=总是|1=经常|2=有时|3=偶尔|4=从不	2	已核实	2020-03-10			
2334	素质/气质	我能在交流前整理好我关注的内容: 0=总是|1=经常|2=有时|3=偶尔|4=从不	3	已核实	2020-03-10			
11	素质/气质	我能在交流时直视对方: 0=总是|1=经常|2=有时|3=偶尔|4=从不	1	已核实	2020-03-11			
11	素质/气质	我能在交流时直视对方: 0=总是|1=经常|2=有时|3=偶尔|4=从不	1	已核实	2020-03-11			
22333	素质/气质	我能在交流时直视对方: 0=总是|1=经常|2=有时|3=偶尔|4=从不	3	已核实	2020-03-11			
								提交确认

三、数据提取与数据下载

数据分析时常常要从不同表单中选择需要的问项（变量），然后合并成一个新的用于统计分析的工作数据文件。为确保系统能正确地将同一研究对象的数据从不同表单里提取并串联起来，课题所用的每个表单在数据开始录入前需预先设置了研究对象编号变量，该变量一旦设定，不应该再更换。这样数据中的每条记录都有一个研究对象的编号（SID）。如有随访（重复测量）且每次重复测量用到多个表单，则还需要预先设置随访（重复测量）次序变量，而且该变量一旦设定，不应该再更换。这样数据中的每次随访（重复测量）记录都有一个研究的对象的编号（SID）与随访次序编号（TIMEID），数据提取时才能通过 SID 与 TIMEID，将同一次随访的信息串联到一起。

数据提取首先选择表单，然后选择表单中的问项（变量），示例如下：

数据提取输出文件包括数据和变量说明文件。

数据下载包括单个表单数据文件下载和一键生成的所有表单压缩（.zip）文件下载。 zip 文件中包括各表单对应的数据、变量说明和表单页面文件，其中数据和变量说明文件均为制表符分隔的文本文件，易倆统计软件可直接读取，表单页面文件为 html 格式文件，方便用户查看。

四、进度报告

一个贴切的进度报告，既有利于研究者掌握现场工作进度，又有利于质量控制。易倆 DataWeb 课题进度报告包括以下内容，充分满足上述要求。

1. 按表单统计的记录数和研究对象数，示例如下：

表单名称	表单编号	记录数	研究对象数	变量分布
营造良好氛围	54	8	8	
保持对话题饶有兴致	55	5	5	
换位思考	58	8	8	
专注与总结	61	3	3	
实验室检查及辅助检查	195	0	0	
读懂讲者	2538	2	2	

通过对比记录数与研究对象数可以提示是否有重复录入，或研究对象编号输入错误导致重复录入情况。

2. 在上图中，每个表单最后一列有“变量分布”按钮，可查看表单内各变量的数据分布，连续性变量的分布参数包括最小值、最大值、常用的百分位数（5%、10%、25%、50%、75%、90%、95%）、均数、标准差、缺失数和频数分布直方图。分类型变量的分布为各取值（分类）的频数和频数分布图。变量分布能确切反映数据质量。

3. 按表单完整性组合统计的研究对象数，示例如下：

营造良好氛围	保持对话题饶有兴致	换位思考	专注与总结	实验室检查及辅助检查	读懂讲者	研究对象数
Y	Y	Y	Y			2
Y	Y	Y				1
Y	Y					2
Y						3

研究人员不仅需要各完成各表单的研究对象数，还需要掌握数据的完整性，即完成所有表单或主要表单的研究对象数。如上图显示完成“营造良好氛围”、“保持对话题饶有兴趣”、“换位思考”、“专注于总结”四个表单，但没有“实验室检查及辅助检查”和“读懂讲者”两表单的研究对象有 2 人。

4. 研究对象编号和各表单记录数一览表，示例如下：

研究对象编号	营造良好氛围	保持对话题饶有兴致	换位思考	专注与总结	实验室检查及辅助检查	读懂讲者
-	1	1	1	1	0	0
22	0	0	0	0	0	1
2444	0	0	0	0	0	1
12345	1	1	0	0	0	0
4333445	1	0	0	0	0	0
2233	1	1	1	0	0	0

通过此表可以查询每个研究对象各表单完成情况，以及是否有重复录入（如记录数大于 1）情况。

易侕大数据处理系统

大数据的一大特点是信息量大和信息多样性。近年来，临床医学公开数据库越来越多，如 MIMIC 重症系列数据库（包括 MIMIC-II，MIMIC-III，MIMIC-IV，eICU，PIC 等），美国 NHANES 数据库，美国 SEER 肿瘤数据库，中国 CHNS 数据库等。对这大型公开数据的开发利用，以及对医院信息系统(Hospital Information System, HIS)数据的开发利用成为热门之一。要开发利用大数据库，首先是要熟悉数据库表单名、变量名和变量分布，在此基础上才可形成合理的科研假设和科研设计，然后根据科研设计从大数据库中提取所需的记录和变量，生成用于统计分析的工作数据文件。因为大数据信息量大，单个表单的记录

数往往超出 Excel 容量范围，通常不能直接用 Excel 进行操作，需通过计算机编程用相应的软件才能实现数据的筛选、提取、合并等操作，而在编程和软件使用时往往因计算机内存容量的限制，增加了编程的难度。这些是目前阻碍大数据开发利用的瓶颈。

易佾大数据处理系统旨在帮助用户突破上述瓶颈，在普通的计算机上（对内存容量不做特别的要求），不需要通过编程就能实现大数据的清理（包括文本信息提取）、筛选记录、提取变量和合并数据等操作，生成用于统计分析的工作数据文件。

从源数据文件清理、筛选记录、提取变量、合并数据文件到最后工作数据文件的形成，往往是一个单向的多步骤操作过程，易佾大数据处理系统自动记录每一步操作过程，可以重演每一个中间文件和最终工作数据文件的生成过程。既有利于对工作数据的生成过程进行核对和质量控制，又有利于数据的更新和重现。特别是当源数据文件有更新的情况下，可通过重复原操作流程自动更新最终用于统计分析的工作数据文件，大大提高工作效率。

本章主要介绍易佾大数据处理的设计思路与基本功能，具体操作详见相应模块的帮助文件和视频。

第一节　数据清理与提取

数据清理中一项基本的操作是清除重复记录，这里所说的重复记录是指错误地输入的两条或多条相同的记录。由于（1）数据库表中可能含有自动递增的指示变量和（2）录入错误，这两个原因导致同一条记录即使错误地录入两次，对应的两条记录也不是所有变量都完全相同。因此在清除重复记录时，要求用户尽可能多地选择变量，如果两条记录所选的变量取值都完全相同，被认为是重复记录，可只保留一条。通常在 Excel 里通过排序找出重复记录不是很难的事，然而当数据量大并有内存容量限制的时候，清除重复记录就不是一件容易的事了。

数据清理中比较复杂的操作是从文本中提取信息。文本有以下两种类型：

（一）清单式的文本，如出院诊断的疾病名，文本书写为：“高血脂、高血压，糖尿病”；曾服用药物名书写为：“补佳乐、阿司匹林、强的松”。这类文本首先要按分隔符分解成若干个名称，然后为每个名称分配一个 0/1 两分类变量。如出院诊断有是否有高血脂？是否有高血压？是否有糖尿病？曾服用药物有是否有补佳乐？是否服用阿司匹林？是否服用强的松？

（二）描述性的文本，如既往病史书写为：“患者性别：男年龄53岁汉族，婚姻状况：已婚血常规：白细胞数 10.66*10^9/L，　中性粒细胞绝对值 9.80*10^9/L,……”。从这类文本中可以提取若干个变量信息，如性别、年龄、婚姻状况、白细胞数、中性粒细胞数等。易佾通过搜索“前文本[所要的信息]后文本”提取相应的数据，前文本与后文本可有可无，给定前、后文本帮助准确查找，如没有前文本与后文本，直接输入所要的信息（省去[]号）。 []内所要的信息是可通过符号，如“#”、“$”、“@”、“+”、“* ”，“?”等来规定信息类型和字符长度，以期达到准确搜索的目的。基本规则如下：

1) # 表示任何数值，包括整数、小数、正数、负数，如 12，0.1，-10.01 等；
2) $ 表示任何字符(含数字)；
3) 可以是 #、$、？、*、+ 的组合，规则如下：
    a) $$: 2 个字符；$$$: 3 个字符，依此类推；
    b) ##: 2 个数字；###: 3 个数字，依此类推；
    c) $?：0 或 1 个字符；$+: 大于等于 1 个字符；$*：大于等于 0 个字符；${1,n}:长度 1-n 之间的字符；
    d) #?：0 或 1 个数字；#+: 大于等于 1 个数字；#*: 大于等于 0 个数字；#{1,n}:长度 1-n 之间的数字；
4) @ 表示任何时间，适用格式如： 2000 年 01 月 01 日 00 时 00 分 01 秒，2000-01-01 00:00:01，2000 年 01 月 01 日，2000-01-01；
5) 所要的信息可以是文字，如“高血压”；
6) 前文本、要查找的信息、后文本可以有多种，中间用 | 分隔，如找电话号码或手机号码打头的后面的数字，可用：电话号码|手机号码[#]；如要找文本中是否有“高血压”或“血压高”的记录，可用：高血压|血压高；
7) 前文本中可用单个 # 代表任何数字，用单个 $ 代表任何一个字符，用 ${n1,n2}代表 n1-n2 位长的任何字符。如果要指定的前文本中含有 # 或/和 $ 字符，要关闭其代替数字/字符的功能，可在前文本前加 ` 符号。
8) 前文本与要查找的信息之间可加 ~~~，代表中间有任何 1-5 长的字符。如：补佳乐~~~[#]天，表示搜索补佳乐+任何 1-5 长度的字符+[数字]天，提取其中的数字；如中间字符可能超过 5 位长，增加~，每增加一个 ~ 表示中间字符的长度增加 5 位，~~~~表示 1-10 位长，~~~~~表示 1-15 位长；
9) 在前文本前加 ` 符合表示不自动对前文本进行任何处理。如前文本中含 # 或/和 $ 字符，默认 # 代表数字，$ 代表字符，要关闭此替代功能，可在前文本前加 ` 符号。另外对于于精通 Javascript 的用户，可直接输入 RegEx 表达式作前文本，这时需要关闭上述的自动替代功能。

如从文本“补佳乐 4mg qd×14 天 黄体酮软胶囊 0.2g tid×14 天 地屈孕酮 10mg tid×14 天”里提取：

1) 补佳乐剂量，用：补佳乐[#]mg，将提取出数字 4。
2) 补佳乐日服用次数，用：补佳乐#mg [${1,2}]d，将提取出字符 q，并提示对 q、bi、ti 三种可能的服用次数进行编码，可分别编为 4、2、3 对应每日服用次数。此时前文本中的 # 代表任何数字，所要的信息 ${1,2} 为 1-2 位长的字符。
3) 补佳乐服用天数，用：补佳乐#mg~~~[#]天，或用：补佳乐~~~~[#]天，将提取出数字 14。前者用 ~~~ 表示在补佳乐后的 mg 与服用天数之间有任意 1-5 个字符，后者表示在补佳乐与服用天数之间有任意 1-10 个字符。

为解决内存限制与速度慢的问题，易倾文本数据清理会首先在一个 2MB 的子数据中运行，用户可以查看子数据的每条原始记录并与清理后的数值进行对照，以确保数据清理过程正确，然后再自动应用到全部数据。

第二节 非结构化数据提取

结构化数据是一个行乘列表的矩阵结构，每一行代表一条观测记录，每一列代表一个观测指标。非结构化数据则不然，不同观测指标的观测值用同一个变量记录在同一列（不同行）中，另有一个变量代表所观测的指标名称。临床电子病例中的实验室检测资料，多为非结构化数据，其基本结构如下图 6-1 示例：

SUBJECT	ITEM	TIME	VALUE	UNIT
3665	5244	10/25/2077 9:23	27.2	mmHg
3665	5247	10/25/2077 9:23	0.86	
3665	5248	10/25/2077 9:23	22.2	mmol/L
3665	5249	10/25/2077 9:23	-0.3	mmol/L
3665	5250	10/25/2077 9:23	61.4	%
3665	5252	10/25/2077 9:23	59.1	%
966	5250	9/20/2068 4:53	36.6	%
966	5249	9/20/2068 4:53	1	mmol/L
966	5248	9/20/2068 4:53	24.7	mmol/L
966	5247	9/20/2068 4:53	0.86	
966	5246	9/20/2068 4:53	15.4	%
966	5245	9/20/2068 4:53	33	mmHg
966	5244	9/20/2068 4:53	51.2	mmHg

非结构化的随访数据的一个特点是同一个病人同一检查指标可能有多条记录，每个病人的检测时间不统一，如何整理这种数据，分析利用这种数据是一个很大的挑战。

使用易倾非结构化数据提取模块提取数据，首先给定检测名称变量（如上图中的 ITEM）与检测结果变量（如上图中的 VALUE），易倾自动列出数据中所有的检测项目，供用户选择。将检测时间转化成相对时间（如随访或入院天数或小时数）后，可以提取：

（1）指定基线与选择随访时间段，输出每个患者每个指标在各时间段的数据。如指定基线时间段为 0-2，表示在 2 天之内检测的结果均计为基线值，如某病人某指标在该时间段内有多次测量，用户可选择使用：第一次测量值、平均值、最小值、最大值、中位值、最后一次测量值。同理，如指定 5-8 为第一次随访时间段，该模块自动将随访天数大于等于 5 并小于 8 的检测结果计为第一次随访值。

（2）只指定基线时间段，置随访时间段为空。输出纵向数据，基线时间段后每次测量即为一条随访记录。

（3）不指定时间段，选择输出每个测量时间点的数据，一个时间点一条记录。

（4）不指定时间段，选择输出所有时间点测量数据的统计量，包括第一次测量值、第一次测量时间、平均测量值、最小测量值、最小测量值的测量时间、最大测量值、最大测量值的测量时间、中位测量值、最后一次测量值、测量次数、所有测量值的标准差、最大上升百分比、最大上升值、最大下降百分比、最大下降值等，可多选。

第三节 大数据记录筛选

根据纳排标准，从大数据中筛选我们需要的记录是必不可少的基本操作。纳排标准往往是由多重标准组成，其选择条件可能来自多个表单，甚至需要联合多个表单的变量进行计算后才能判断。如“年龄大于等于 50 岁，性别为男性，在重症监护病房停留至少 2 天”，其中年龄和性别存在于患者的一般情况表中，而在重症监护病房（ICU）停留时间则需要通过计算出 ICU 与入 ICU 的时间差，而且这两个时间变量又可能存在于不同的表中，这就给大数据记录筛选带来了复杂性。

易侕大数据记录筛选模块自身带有简单的创建新变量功能，这样如果筛选条件需要通过简单的变量计算来实现，如计算前后时间差，可以轻松实现。页面图 6-2 如下：

筛选条件示例截图如下：

如上图 6-3 所示，筛选记录的途径有：

根据筛选条件，从单个表单中筛选记录，如 LOS >= 2。多重条件可以通过 AND 和 OR 及括号规定优先顺序。

从重复 ID 记录(如重复测量或多次入院等)里筛选第一条记录或最后一条记录，当然这里所谓的第一条和最后一条是按某变量排序后来确定的。

匹配另一文件中的 ID，设想场景：一个大数据含表单 1：病人基本情况(包括疾病诊断)，表单 2：生化检测记录，表单 3：随访记录，等等，根据表单 1 中的病人编码从表单 2、表单 3..... 中提取记录。勾选此项并选择匹配文件（如表单 1），选择条件是当前文件中 ID 变量名与匹配文件中的 ID 变量名相等，研究对象编码可以是多个变量组成的梯次编码，如医院编码、科室编码、个人编码。此时，筛选条件除了指定变量取值范围(如 AGE > 50)外，还可以将来此两个文件的变量进行比较，如下图 6-4 显示条件有当前文件中 CHARTTIME 大于等于匹配文件中的 INTIME，小于等于 OUTTIME。

第四节 匹配 ID 导入变量和大数据合并

数据库中研究对象的信息存放在不同的表中，根据研究对象编码（ID）将几个表（数据文件）的变量横向合并，也就是将研究对象的信息串联起来，生成一个工作数据文件，这个过程看上去简单，实际操作起来有很多的陷阱，要做到准确无误，需要首先对每个表的编码和变量进行详细检查，确定对下述问题清楚掌握了后才进行合并：

1. 多个数据文件中变量名是否有交叉重叠？如有两个或多个数据文件有同名变量，其变量分布是否一致？检查在不同文件中的同名变量的分布，有利于帮助我们判断同名变量是否是同一观测指标。

2. 各数据文件是否有同一表示研究对象编码的变量？如没有，则不能一次性将多个文件合并。如有，各数据文件的研究对象编码是否有重复？如有重复而且不是重复测量数据，则说明编码重复是由于录入错误导致的，首先需要清错，然后才能合并。来自这些要合并的数据文件的研究对象编码交叉重叠情况如何？掌握了这个情况，才能判断最终合并起来的数据有多少记录是完整的。

易侕大数据合并模块包括数据的纵向合并（首尾相连添加记录）和横向合并（添加变量）。纵向合并要求各数据变量名相同。横向合并在读取数据后，首先给出各数据文件的变量名和变量分布，下图是要合并 regis1.csv、ques1.csv 和 labg1.csv 三个文件，检查变量示例结果截图 6-5：

检查变量 连续性变量 分类型变量 含非数字字符 全为空 [取值数 / 缺失数] 选择输出变量

勾选ID变量	变量名	regis1.csv (N=436) 全选:	ques1.csv (N=428) 全选:	labg1.csv (N=429) 全选:
	SUBJ	[428/0]	[427/0]	[428/0]
	NID	[9/0]		
	SEX	[2/0]		
	FMYTYPE	[2/0]		
	AGE	[281/0]		
	FMYID	[102/0]		
	OCCU		[10/0]	
	EDU		[4/0]	

上图中“钟形”绿色图标表示连续性变量，“三竖条”绿色图标表示分类型变量，点击图标可以进一步看到变量的具体分布。检查结果通过列表方式显示，方便查看变量名交叉重叠情况。

勾选研究对象编码（ID）变量后，检查 ID 即可得到研究对象编码在各数据文件的分布情况，包括是否有重复的编码和编码的交叉重叠情况，并可选择最终要输出的编码。示例图如下：

图 6-6 中，对每个数据文件列出 ID 总数/ID 取值数，ID 总数是有 ID 的记录数，ID 取值数是有多少种不同的取值，如果取值数小于总数表示有 ID 重复使用，同一个 ID 有两条或多条记录。点击“清单”的图标显示重复的 ID。

数据合并时对重复 ID 的处理通常有三种选择，一是所有记录都需要输出（如重复测量的数据文件，输出每次测量的数据），二是只取第一条记录，三是只取最后一条记录。数据合并时只能允许其中有一个数据文件有重复 ID 记录输出。如果 ID=XXX 的人在文件 1 中有二条记录，在文件 2 中三条记录，在文件 3 中有 4 条件记录，如果将这些记录串联，就有 2*3*4=24 种可能的组合，显然是不合理的。

如果是重复测量数据，同一次测量的数据存放在多个文件中，应该通过研究对象编码与测量次序两个变量将同一次测量的结果串起来，而不能取组合。此时应指定 ID 变量为研究对象编号与测量序号两个变量。易倾数据合并模块可以接受复合 ID 变量，即由多个变量组成的复合 ID。

检查 ID 输出的结果给出了 ID 在各文件中的组合情况，如上例中有 428 个 ID 只出现在 labg1.csv 文件中，427 个 ID 出现在前 2 个文件（regis1.csv 与 ques1.csv）中。最后 1 列在 ID 数后面有一个表示“清单”的图标，点击该图标查看这些 ID，如点击图例中 428 个只出现在 labg1.csv 中的 ID 图标，查看这些 ID 取值可以帮助我们了解什么原因导致三个文件没有完全重叠的 ID。如上图例中，ID 在 regis1.csv 和 ques1.csv 中为固定 5 位长的数字字符如 00001、00012，而在 labg1.csv 中为 1、12 纯数字，需要先将它们统一，然后才能合并。

当没有相同变量名表示研究对象编码时，可采用“匹配 ID 导入变量”模块，该模块是将另外一个文件中的变量导入到当前数据文件中，相当于合并两个文件。实际上多个数据的合并也可以分解为多个“匹配 ID 导入变量”操作。使用该模块，ID 变量也可以是多个变量组成的复合 ID，匹配数据中的 ID 变量名与当前数据文件的 ID 变量可以不同名。示例截图如下：

第五节 数据转换

每个统计软件都有自己的数据文件格式，如 SAS 的数据文件是.sas7bdat 文件，SPSS 的数据文件是.sav 文件，R 的数据文件是.Rdata，EpiData 的数据文件是.rec 文件。这些软件一般都不能直接读取其它软件的数据文件，但应该都可以读取文本文件数据，而且都可以将其本身的数据文件输出成一个文本格式的数据文件。一般来说，通用的数据文件格式是制表符分隔的文本文件，任何统计软件都可以读入。制表符被认为是最佳分隔符，因为字段取值如果是长串的文字描述，里面很可能需要有逗号或/和空格，但一般都不会需要制表符。用制表符分隔的文本文件一般不会导致错位和串行。

如果文本中含有中文，中文在计算机内的编码最佳选择是用 UTF-8 编码。UTF-8 使用 3 个字节表示 1 个汉字，GB2312 编码则使用 2 个字节。如果文件内有很多汉字，ANSI / GB2312 会节省空间，但如果只是英文字母和数字，UTF-8 和 ANSI 占用空间完全相同。然而，UTF-8 可以节省将来的麻烦，转换为 UTF-8 后，无需担心字符集或编码。UTF-8 在国际上更具字符友好性，大多数浏览器都知道如何正确显示 UTF-8 文本。

易侕数据转换模块可将如下三类文件转换成 UTF-8 编码的制表符分隔的文本文件：

（1） Excel 后缀为.xlsx 或.xls 的文件
（2） EpiData 后缀为.rec 的文件
（3） 无分隔符的固定字段长度的.dat 文件

Excel 的.xlsx 或.xls 文件，如用 Excel 打开，在 Excel 环境下也可直接保存成制表符分隔的文本文件，然而当文件内含中文时，出来的文件中文字符通常不是 UTF-8 编码。用易侕数据转换模块，自动规避了 Excel 中文编码的问题。

无分隔符的固定字段长度的.dat 文件没有字段名，第一行开始即是数据，字段之间没有分隔符，每个字段实际上有其固定的长度。如一行数据为 0000700552223202032。要读取该文件首先需要知道数据有哪些字段，每个字段的长度，然后才能依次将每行数据分解为字段数据。这就需要有一个字段说明文件注明字段名与数据在每行中的位置。如：SEQN 1-5 DMPF 6-10 STAT 11 THN 12，表示第 1-5 字符是字段 SEQN，6-10 字符是字段 DMPF，第 11 字符是字段 STAT，第 12 字符是字段 THN。

如上所述，无分隔符的固定字段长度的.dat 文件没有字段名与分隔符，只是数据，所以需要有字段说明文件指导才能读取。EpiData 的 REC 文件相当于此类文件但把字段说明放到数据之前合成一个文件。REC 文件的第一行记录了有多少个字段，第 2 行开始是每个字段的字段名、类型、数据位置，每个字段一行，然后是数据行。REC 文件的数据行每行的长度固定为 80 个字符，如一条记录长度超过 80 个字符将分成多行存放。

易佾软件直接读取制表符、逗号或空格分隔的文本文件。文本文件的大小不受软件限制，从大型数据库上下载的数据文件通常存成制表符或逗号分隔的文本文件。易佾大数据处理模块直接读取制表符或逗号分隔的文本文件。

（陈常中）