·AME 科研时间专栏·

专栏导读：AME Groups 旗下出版了 *Journal of Thoracic Disease*（《胸部疾病杂志》）、*Annals of Cardiothoracic Surgery*（《心胸外科年鉴》）、*Chinese Journal of Cancer Research*（《中国癌症研究》）和 *Annals of Translational Medicine*（《转化医学年鉴》）等近 20 本英文医学学术期刊。2014 年，AME Groups 中文平台——“科研时间”的诞生，为广大从事临床和基础研究的科研工作者带来了福音，提供了更多科研交流和学习分享的机会。欢迎广大读者关注我们“AME 科研时间专栏”，订阅我们的公众微信号（科研时间：amegroups），给我们提出宝贵的建议和意见，以便于将这个专栏建设得更好，成为读者喜闻乐见的一个栏目。

doi: 10.3978/j.issn.2095-6959.2014.05.009
http://www.lcbl.net/articles/639

大数据与临床科研

章仲恒

(金华市中心医院重症医学科，浙江 金华 321000)

[摘　要]　21世纪是信息时代，大数据概念包围着人们生活的方方面面。在医疗科研领域正面临着一次前所未有的变革，电子病历信息系统的发展及其他各种医疗数据的信息化存储将引领临床科研工作者进入这个充满机遇与挑战的时代。本文将根据笔者的实践经验对大数据在临床科研领域的应用进行探讨。主要从临床科研面临的困境、大数据实例简介、科研思路的凝集、数据处理等几个方面进行探讨。希望本文能为有兴趣致力于研究大数据的临床科研工作者提供一点思路，起到抛砖引玉的作用。

[关键词]　大数据；临床科研；电子病历；医药卫生

Big data and clinical research

ZHANG Zhongheng

(Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Zhejiang 321000, China)

Abstract　The 21st century is an era of big data involving all aspects of human life including economics, politics and healthcare. A big-data revolution is under way in health care, starting with the vastly increased supply of electronic medical record system and other healthcare databases. The narrative review will discuss something on the use of big data for clinical researches, by incorporating the author's experience. The discussion will include several aspects such as limitations of conventional clinical research, the example of big data, conceiving ideas of clinical research

收稿日期 (Date of reception)：2014–08–18
通信作者 (Corresponding author)：章仲恒，Email: zh_zhang1984@hotmail.com

and data management. I hope this article will provide new insights into the use of big data for clinical investigators and solicit more brilliant work on this field.

Key words big data; clinical research; electronic medical record; healthcare

1 前言

21世纪人类全面进入了电子信息时代，这是一个革命性的进步，从此人们步入了知识信息大爆炸时代。海量的数据信息有助于人们更加客观的认识和掌握各种自然及社会规律，维基百科对大数据做出了一个定义：“大数据指的是所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息”。2012年美国的时代周刊对大数据带来的社会变革也提出了“大数据时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉”。在健康卫生领域，大数据同样能发挥其巨大作用，对于病人的诊疗策略更加需要基于“数据分析”而得出，而非传统的经验和直觉[1]。

本文将探讨如何利用大数据来进行临床科研。文章首先简单介绍临床研究的一些基本知识以及医学大数据的一些特征；随后将通过一个实例并结合笔者已经发表的研究论文来介绍如何利用大数据进行相关临床问题的探讨。希望本文能给有志于研究大数据的临床工作者带来一定的启发，起到抛砖引玉的作用。因为数据分析是一项实践性极强的工作,许多具体的细节无法一一列出，为此我附上了许多参考文献，感兴趣的读者可以查阅有关文献来进一步探索大数据。

2 当前临床研究所面临的困惑

众所周知，临床研究大体而言可以分为干预性研究和观察性研究(为求简化，我们暂不讨论类试验研究及有干预无对照研究)，前者属于一种试验性研究，临床上多称为“随机对照临床试验(randomized controlled trial，RCT)”需要主动地、有研究目的地对受试者进行试验干预，一般有严格的纳入排除标准，采用随机化的方法来最大程度地消除混杂因素的干扰，这也是当今循证医学的“金标准”，各类指南以及高级别循证医学证据均来源于这类研究或者相关的荟萃分析。但随着大量RCT的开展，其弊端也日益暴露出来。在脓毒症及重症医学领域，我们曾经对观察性研究和RCT得出的结论进行比较，结果发现两者差别较大，也就是说两者干预效应(interventional effect)并不一致[2-3]。RCT得出的是一种生物学疗效(biological efficacy)，这反应了干预手段在严格的实验条件下的生物学作用。而事实上，这种作用可能很弱或者在实际的临床工作中无法发挥出来，及生物学效应不能转化为临床疗效(clinical effectiveness)[4]。而临床疗效其实是我们临床医生最为关心的问题。至于为什么有生物学疗效的东西不能转化为临床疗效呢？这主要是因为RCT条件下干预措施执行比较严格，其中包括纳入的人群(没有大量的合并症和并发症，是所谓的单病种，这就无形中剔除了大量的混杂因素)以及严格的治疗时机(有时候现实中繁忙的临床工作可能会延误给药时机)。另外RCT往往在一些大型的医疗机构进行，其结果并不能推广到一般的中小型医院。同时也因为RCT严格的纳入标准(有兴趣的读者可以去看一些RCT原文，都可以发现一长串的所谓的排除标准)，导致目前的循证医学存在这样一种现象，即利用从20%患者身上得出的结论去治疗其余80%的患者。另外观察性研究也存在选择偏倚、混杂因素难以控制、基线资料不均衡等各种缺点。

目前尚没有一种可靠的办法来解决这个问题，但大数据似乎能给我们一些新的启示[5-6]。有著名学者曾指出基于大数据的临床研究(BCT)在未来可能成为一种研究的主流方式并与RCT形成互补[7]。医疗卫生领域的大数据是指因为临床或者科研需要收集起来的有关健康或者诊疗的信息。其产生都来源于日常诊疗过程，其中的数据都是未加修饰的“纯天然”数据，也就是说，里面没有任何的排除纳入标准，而且可以认为是一种普查数据(census data)。大数据所能直接反应的就是临床疗效[8]。卫生领域的大数据大致可以包括医院的电子病历系统、慢性病及传染病注册登记系统、医疗保险信息系统、出生死亡登记系统等等[9]。当然不同的国家和地区以及对医疗卫生不同的管理方式会产生各种不同的大数据信息。作为临床医生我们最为关注的还是电子病历信息系统[10]。该系统包括了人口学特征(性别、年龄)、实验室检查、微生物检测、医嘱、诊疗操作、手术资料和临床转归等信息。然而大数据并不是万能的，从本质上来说，大数据其实属于观察性研究，因此必然存在观察性研究的缺陷，例如存在较

多偏倚、基线资料难以均衡以及混杂因素难以控制等。如果分析了存在偏倚的数据，得出的结论有可能被夸大，目前尚不存在一种万能的统计学解决方案，但似乎可以通过采用随机效应模型或者bootstrap抽样法在一定程度上提高预测模型的准确度。

3 如何利用大数据进行临床研究：以MIMIC-II为例

该部分内容将采用笔者较为熟悉的MIMIC-II数据库作为实例进行讲解[11]，其大致流程见图1。

MIMIC-II数据库全称为重症监护多参数智能监测数据库II(Multiparameter Intelligent Monitoring in Intensive care II database)。该数据库是对公众开放的免费数据库，主要用于重症医学的各种临床研究，其网址为http://physionet.org/mimic2/。MIMIC-II里面包含的患者信息来自于美国波士顿贝斯以色列女执事医疗中心(Beth Israel Deaconess Medical Center)。该数据库是不断更新的，目前使用的是2.6版，收集了2001年到2008年三万多个ICU患者的住院信息。MIMIC-II包含的信息有人口学特征、实验室检查、液体及药物医嘱、病历信息和护理记录。另外一大块内容包括高精度的波形记录，包括心电监护、呼吸波形监护、血压、指测血氧饱和度等，这类数据的获得与危重症病人密切监护分不开[12]。

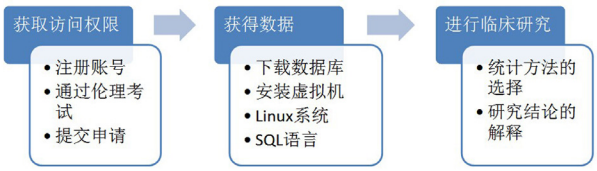

图 1 MIMIC-II 使用流程图

Figure 1 flow chart of using MIMIC-II

4 数据库的获得

用户首先要在网站申请一个用户名，申请成功后提出访问数据库的申请。其中需要通过一个有关临床研究伦理学的学习考试，随后就能获得一个由美国国立卫生院(national institute of health，NIH)颁发的证书(图2)，上面同时会给出一个证书编号，凭该编号就可以去MIMIC-II申请访问权限了。申请成功后你就可以对全部数据进行下载，根据自己感兴趣的研究内容展开相关的数据分析。


Certificate of Completion

The National Institutes of Health (NIH) Office of Extramural Research certifies that **Zhongheng Zhang** successfully completed the NIH Web-based training course "Protecting Human Research Participants".

Date of completion: 02/28/2013

Certification Number: 1132877

图 2 美国国立卫生院 (national institute of health，NIH) 颁发的完成“保护临床受试者”课程的结业证书

Figure 2 completion certificate of protecting human research participants provided by the national institute of health (NIH)

5 科研思路的凝集

当我们下载了数据之后，下一个问题就是讨论如何利用这浩瀚的数据进行临床研究。首先需要阐明的就是利用大数据可以进行的几种研究(表1)。首先就是危险因素的评估，此类研究往往需要比较高密度的临床资料用于混杂因素的控制，研究方法上可采用多因素回归分析,分层分析以及倾向性评分。第二种是干预手段的效果评价，这种研究需要高精度的临床数据来控制混杂因素，并探讨

干预手段是否因为“选择性治疗”而有所不同。第三种是预测模型的建立，该类研究与危险因素研究最大的差别就是其所关注的不是某个危险因素，而是整个模型的预测价值。第四种是流行病学的研究，该研究仅仅采用简单的描述性研究，不需要危险因素矫正及组间比较。最后一种是评价某种医疗政策和手段实施的效果，常常不需要很复杂的临床指标，因此临床数据要求比较低。

如何凝练出实际可行的临床研究问题又是另外一个重要环节。这大概可以分为两种方式：一种从数据找灵感，另外一种就是从要研究的内容去找数据。这两种方式看似相反实则统一。有时可以相互结合着用。比如我有一个想法，希望研究初始乳酸与ICU患者预后的相关性，结果我进行数据发掘(data mining)时候发现每个病人查了很多乳酸值，而不是一个，有的甚至是每隔几小时就查一个。于是我就会在原有设计基础上进行调整，比如把初始乳酸改为6小时乳酸清除率，这样就更进一步了。这就是说原有的设计思路可以根据具体的数据进行细微的调整。

从数据找灵感也是一种方式，拿到一份数据资料后首先对变量进行简单的描述性统计分析，包括了各个变量的集中趋势、离散趋势、全距以及分布类型。笔者喜欢通过作图来查看这些数据，比如轮廓图(contour plot)可以查明各种数据之间是否存在一定的关系；又比如用柱状图可以查看数据的分布类型。其他还有很多方法来“阅读”数据，这个“阅读”其实并不是想象中的那样枯燥乏味。当然有人可能会反对我这种做法，因为这种事先了解数据特征再进行研究在统计学中是不行的，犯了多重检验(multiple testing)的错误。也就是说我先检验，然后拿阳性的结果进行报道，那么我进行20次检验就很可能会有一次阳性(P=0.05)，虽然这个阳性是随机效应造成的。我目前无法解释这个问题，但人的天性让我们很想先看看数据是什么再去做研究——因为现实太残酷，你所想的点子未必有数据让你研究，也就是所谓的“巧妇难为无米之炊”。

当然还有一种方法是使用简单的指标，这样就不怕没有数据可以研究了。比如我之前做的关于尿量与危重症患者病死率的关系的研究[13]。因为尿量是ICU必须记录的一个指标，它太重要了而且记录很简单。也没有任何理由说哪个医保政策会限制记录尿量，就像人的吃饭都是三餐的，在美国也好、欧洲也好，这种人类共性的东西我们可以很有信心地说它们一定在数据库中。因此尿量的研究就进行得很顺利，到头来整理的数据也都是满满的一大堆。诸如此类的还有一些简单的实验室指标如电解质、血常规以及体温，我做的关于离子钙方面的研究就是一个例证[14]。当然这种简单的临床指标虽然容易施行，但往往缺乏创新性，这是我投稿中遇到的最大问题。

表 1 利用大数据开展的研究类型列举

Table 1 Classification of clinical researches by using big data

研究类别	举例子（研究问题）	临床资料要求[†]	备注
危险因素的评估（独立性）	入院尿量的多少是否与 ICU 病死率独立相关?	高密度（对该危险因素的细致描述）	可采用多因素回归分析，分层分析以及倾向性评分的方法进行。
干预手段的效果评价	使用 PiCCO 进行血流动力学监护是否能提高脓毒症休克患者的预后?	高密度（包括大量的混杂因素）	许多干预手段是基于患者的不同情况而采取的，而这种所谓的“不同”就是需要控制的混杂因素
预测模型的建立（预测公式）	如何进行 ICU 患者谵妄发生的预测?	中等密度（对各个危险因素有一般的描述）	所关注的不是某个危险因素，而是整个模型的预测价值。
流行病学研究	ICU 病房导管相关的血行感染发生率是多少?	低密度	流行病学研究仅仅采用简单的描述性研究，不需要危险因素矫正及组间比较。
医疗政策和手段实施的效果	高血压防控（普查及治疗）的政策是否降低心血管事件的发生率?	低密度	常常不需要很复杂的临床指标。

[†] 临床资料的密度是指数据记录的密度，比如研究尿量需要记录每小时尿量、如果只记录了每日尿量，我们就说该资料密度不够。又比如要研究各种混杂因素，需要各种相关的协变量，如果混杂因素提供不足我们就认为该数据库密度不足。

6 数据的提取

MIMIC-II数据库分为网络上的IE版本和完全版本，IE版本是给用户体验用的(https://mimic2app.csail.mit.edu/querybuilder/)，能够进行SQL语句的测试，但不能下载到全部的数据信息，这部分可以作为正式开展研究前的预实验使用，比如你可以查看该数据库是否包含你说要关注的某项实验室检查。在开始接触MIMIC-II的时候笔者正在做有关脑钠肽及其前体方面的研究[15-16]，于是我很自然想到是不是可以用该数据的大样本优势进行进一步研究呢？开始我信心满满以为可以拿出一项非常漂亮的研究，于是开始了艰难的数据库下载和提取过程，结果到最后发现，该数据库里包含的BNP信息非常少，我才恍然大悟也许美国的医保并不允许没有适应症的病人去检测BNP，前面花费的功夫就算白费了，其实在处理完全版本的数据库之前完全可以先用IE版本对数据进行预查看，这样如果不可行的研究就可以及早终止，避免时间和精力的浪费。IE版本的提取并不困难，只需在Windows操作系统下就可以进行了，这方面就不再赘述。

完全版本的提取对初学者稍微有点难度，需要有关虚拟机以及Linux操作系的一些知识。在网站上首先下载到的是“.tar”格式的压缩包，大约30 G的大小，其实这是可以直接再虚拟机上打开的，我使用的虚拟机是甲骨文公司的virtual box，直接打开后输入用户名密码就直接进入了linux系统，然后就像登陆windows一样登陆该系统，其用户名和密码分别为“mimic2”和“2CIMIM_2v6”，这是系统默认的不建议修改。进入系统后运行pgAdmin软件，这是调用数据库的，最终使用SQL语言进行数据信息的提取。SQL语言也是关键留到下一个部分讨论。

数据提取之后需要将数据库导出到windows下进行分析，当然如果有条件也可以直接在linux系统下分析。转出来的数据是“.cvs”格式的，可将其直接导入Stata软件，或者用excel转化之后再导入Stata。本人没有用过SPSS，所以只有大家自己去实践体会了。

7 STATA 模块数据调用与分析

笔者在数据分析时主要使用的是Stata统计软件包，因此本文就Stata进行的数据处理及统计分析做一概述。

使用Stata进行大数据分析时主要包括两个部分，即数据处理(data management)和统计分析(statistical analysis)。根据个人经验，大约80%时间和经历将花费在了数据处理上。数据处理包括了以下一些方面：1)新变量的产生，比如希望把年龄变量转化成二分类变量(以65岁以上为老年)；2)数据核查，比如采用sum模块查看是否存在一些不合逻辑的错误值(如年龄变量=200)；3)变量类型的调整，比如有些字符变量转化成数据变量；4)数据库的合并，患者的不同信息往往是存储在不同的关联表格中，比如MIMIC-2数据库中化验检查和医嘱就分别存放在两张表格中，这时就需要进行合并，可以使用stata的merge或append命令执行。统计分析在数据处理的基础上进行，一个完好的数据处理前期准备是进行可靠统计分析的基础。

利用Stata进行分析的一个优点在于能全程记录下对初始数据进行的调整。Stata数据分析可采用三种方式进行，即窗口界面、命令栏输入命令、和do-file。笔者认为任何数据处理都必须采用do-file程序进行，因为do-file可以完整记录数据分析的全过程，这样在文章修回或者发现统计结果错误时可以随时查看调用的命令，从而很方便地解决问题；另外对数据的分析通过do-file传递可以在不同研究者之间进行交流重复。而窗口界面和命令栏输入命令的数据处理方式主要用于stata语句的调试。

一个完整的大数据分析流程图见图3，笔者建议将整个分析工作分为数据处理和统计分析两条线路进行，左边是数据处理，会改变内存中的数据库，因此也就产生了许多以“.dta”为后缀的新数据库表格，而右边仅仅进行统计分析，不改变数据库，因此只有do-file。

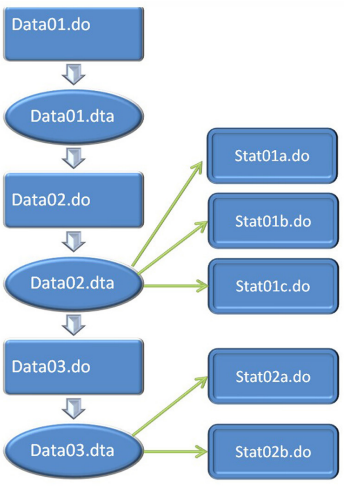

图 3 数据处理与统计分析双线流程图

Figure 3 Dual flowchart of data management and statistical analysis

8 SQL 语言的一点经验

数据导出的一个关键点就是如何使用SQL语句来提取自己想要的数据，这样可以大大提高效率节省硬盘。有人说，我可以把全部数据提取出来之后再用统计软件进行数据处理(data management)，这当然可以，但这体现的是你对统计软件的功力，而且往往需要重新设置统计软件的内存了。其实我本人就是喜欢用后一种办法，但这里我想讨论数据库本身，而不是讨论具体统计软件的使用，所以我就班门弄斧讨论一下SQL语句的一些简单使用方法。

SQL都是如下格式“select(变量名)from(表格名称)where(条件设定)”，其中变量名可以是一个关联表格里的各种变量，如果要选取所有变量此处用“*”，表格名称就是指关联表格的名称，而条件设定比如限定年龄在65岁以上就用“age>65”。这是基本的语句结构，如果有兴趣的读者要学习更加复杂的，更多嵌套结构的可以参照相关的书籍。如张权编写的《SQL查询的艺术》、日本的MICK编写的《SQL基础教程》等都是很好的参考书。这里因为篇幅关系还有笔者对SQL语言的肤浅认识，具体内容就不再赘述了。

9 结语

数据库的探索研究是一个尝试与失败(trial and error)的过程,有人说这是一个令人沮丧的工作，但我觉得这艰难的旅程时刻充满着成功与惊喜，人体与疾病的奥秘也许就隐藏在这浩瀚的数据之中，期待我们去破译和理解。希望本文能激发临床工作者对大数据探索的热情，并建立我们自己的高质量大数据库，为人类的健康事业贡献一份力量。

参考文献

1. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data[J]. J Am Med Inform Assoc, 2014. [Epub ahead of print].
2. Zhang Z, Ni H, Xu X. Do the observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis[J]? J Crit Care, 2014. [Epub ahead of print].
3. Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine[J]. J Clin Epidemiol, 2014, 67(8): 932-939.
4. Nallamothu BK, Hayward RA, Bates ER. Beyond the randomized clinical trial: the role of effectiveness studies in evaluating cardiovascular therapies[J]. Circulation, 2008, 118(12): 1294-1303.
5. Schneeweiss S. Learning from big health care data[J]. N Engl J Med, 2014, 370(23): 2161-2163.
6. Psaty BM, Breckenridge AM. Mini-Sentinel and regulatory science--big data rendered fit and functional[J]. N Engl J Med, 2014, 370(23): 2165-2167.
7. Wang SD. Opportunities and challenges of clinical research in the big-data era: from RCT to BCT[J]. J Thorac Dis, 2013, 5(6): 721-723.
8. Albert RK. “Lies, damned lies ...” and observational studies in comparative effectiveness research[J]. Am J Respir Crit Care Med, 2013, 187(11): 1173-1177.
9. Cooke CR, Iwashyna TJ. Using existing data to address important clinical questions in critical care[J]. Crit Care Med, 2013, 41(3): 886-896.
10. Peters SG, Buntrock JD. Big data and the electronic health record[J]. J Ambul Care Manage, 2014, 37(3): 206-210.
11. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database[J]. Crit Care Med, 2011, 39(5): 952-960.
12. Scott DJ, Lee J, Silva I, et al. Accessing the public MIMIC-II intensive care relational database for clinical research[J]. BMC Med Inform Decis Mak, 2013, 13: 9.
13. Zhang Z, Xu X, Ni H, Deng H. Urine output on ICU entry is associated with hospital mortality in unselected critically ill patients[J]. J Nephrol, 2014, 27(1): 65-71.
14. Zhang Z, Xu X, Ni H, et al. Predictive value of ionized calcium in critically ill patients: an analysis of a large clinical database MIMIC II[J]. PLoS One, 2014, 9(4): e95204.
15. Zhang Z, Zhang Z, Xue Y, et al. Prognostic value of B-type natriuretic peptide (BNP) and its potential role in guiding fluid therapy in critically ill septic patients[J]. Scand J Trauma Resusc Emerg Med, 2012, 20: 86.
16. Zhang Z, Ni H, Lu B, et al. Changes in brain natriuretic peptide are correlated with changes in global end-diastolic volume index[J]. J Thorac Dis, 2013, 5(2): 156-160.

本文引用： 章仲恒 . 大数据与临床科研 [J]. 临床与病理杂志，2014, 34(5): 492-497. doi: 10.3978/j.issn.2095-6959.2014.05.009

Cite this article as: ZHANG Zhongheng. Big data and clinical research[J]. Journal of Clinical and Pathological Research, 2014, 34(5): 492-497. doi: 10.3978/j.issn.2095-6959.2014.05.009