第四届中国R语言会议
华东师大逸夫楼，上海2011

# R语言在新药开发中的应用

*一个“老”临床药理工作者对数据分析软件的选择*

*Jun Shi（史军）, MD, MSc, FCP*
*Clinical Pharmacology, Roche pRED China*
*Nov 12th, 2011, Shanghai, China*
*junshi88@yahoo.com*

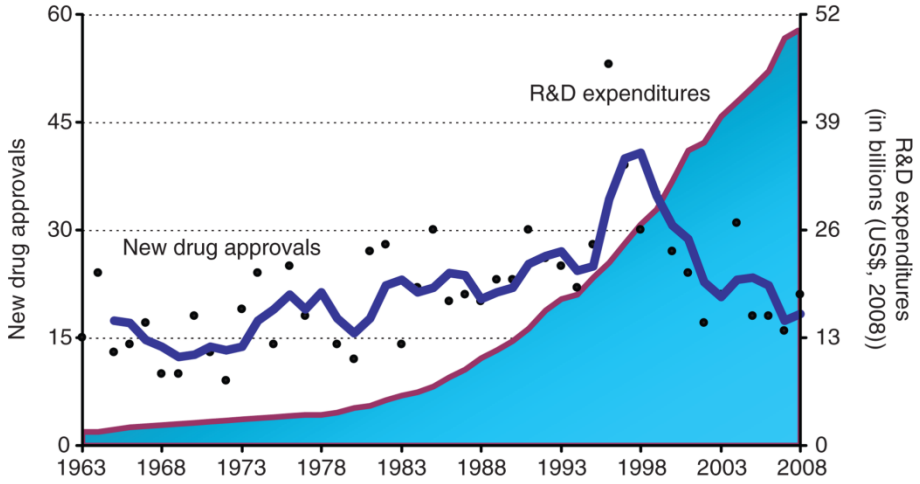
# 提要(outline)

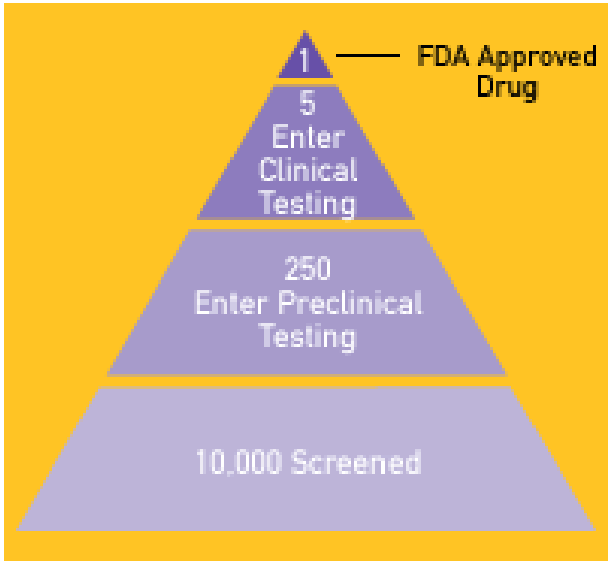
- 个人经历与新药开发的产业（国际与国内）现状
  - 新药开发的瓶颈以及打破瓶颈的方式（Innovation vs Stagnation）
  - “科学联姻”，行业互补和技术互动 （Cutting Edge Science）
- 如何选择数据分析软件 （Software Selection）
  - 从新药开发的视角
- R语言在新药开发中的应用实例（Case Studies）

# 新药开发的特点和现状Pharma R&D Landscape

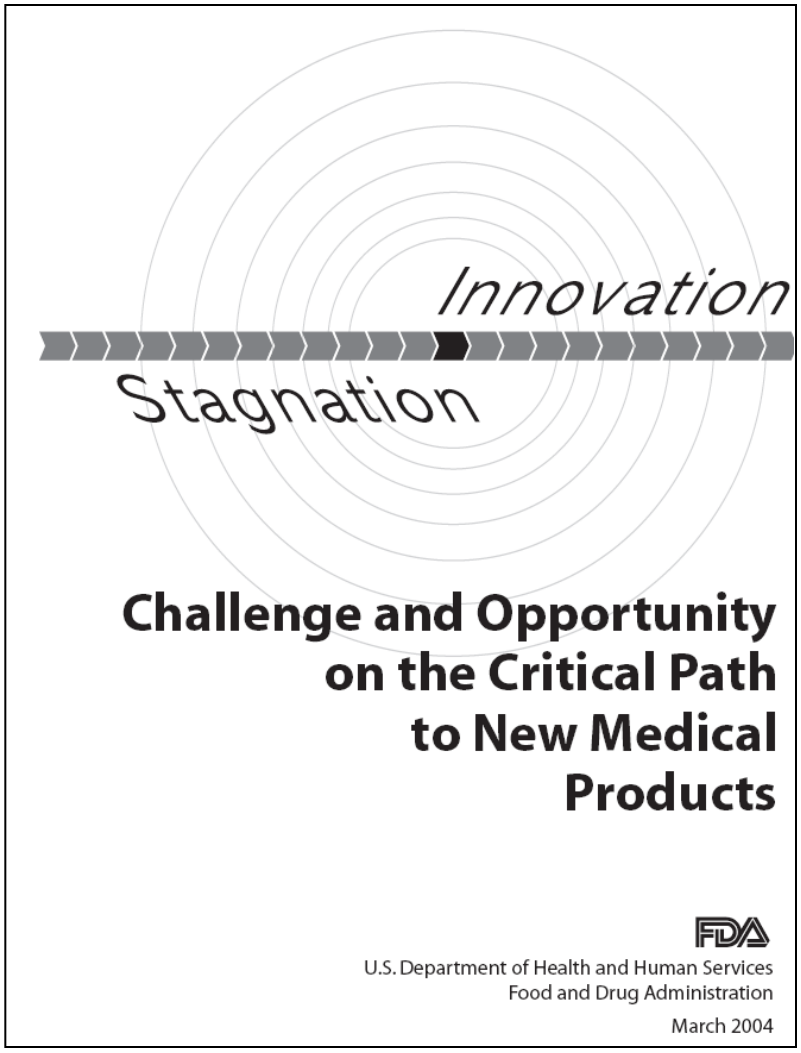
- 新药开发是高成本、高风险，研发周期长，但一旦成功即转化为高利润的产业
- 新药开发是生命科学应用前沿，处在政府和社会的严格监管之下
- 新药开发难度高，在异质性的人群中求得安全性和有效性的统一，以及益损比的合理性
- 现阶段，只有美欧日具有新药开发的经验，但均遭遇前所未有的困境
- 目前，中国还没有完整的新药开发产业链—但作为新兴市场的龙头，拥有众多的未治疗病人群和具有创新精神的科技人员，中国的新药开发前景可观

*Tufts CSDD 2008*



# 美国ＦＤＡ关键路径白皮书
# FDA Critical Path Initiative (2004)



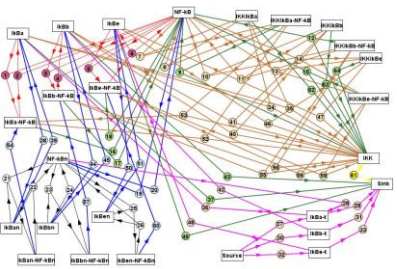

Application of New Scientific Knowledge to Drug Development

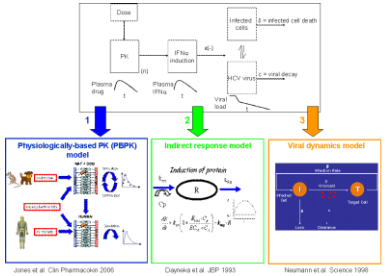
- Application of ***quantitative disease models*** to drug development
- Pharmacogenomics in drug development
- New imaging technologies may contribute biomarkers in drug development

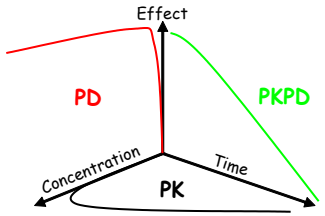
http://www.fda.gov/oc/initiatives/criticalpath/

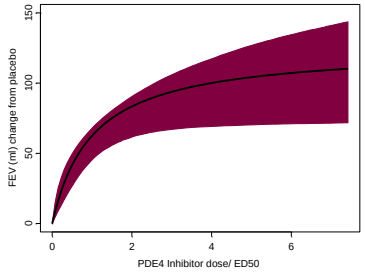
# 基于模型的新药开发（Model Based Drug Development)模式
## Learn-Confirm Cycles连续反复地建模模拟预测验证的过程

| Pathway | Target | Drug & pharmacology | Benefit/Risk | Effectiveness & Reimbursement |
|---|---|---|---|---|






$

| Systems Biology | Systems Pharmacology | Preclinical PK-PD | Clinical PK-PD & disease modeling | Outcome & commercial M&S |
|---|---|---|---|---|

***'Right target'*** ***'Right molecule'*** ***'Right trial & dose'***

***'Right pathway'*** ***'Right patients'***

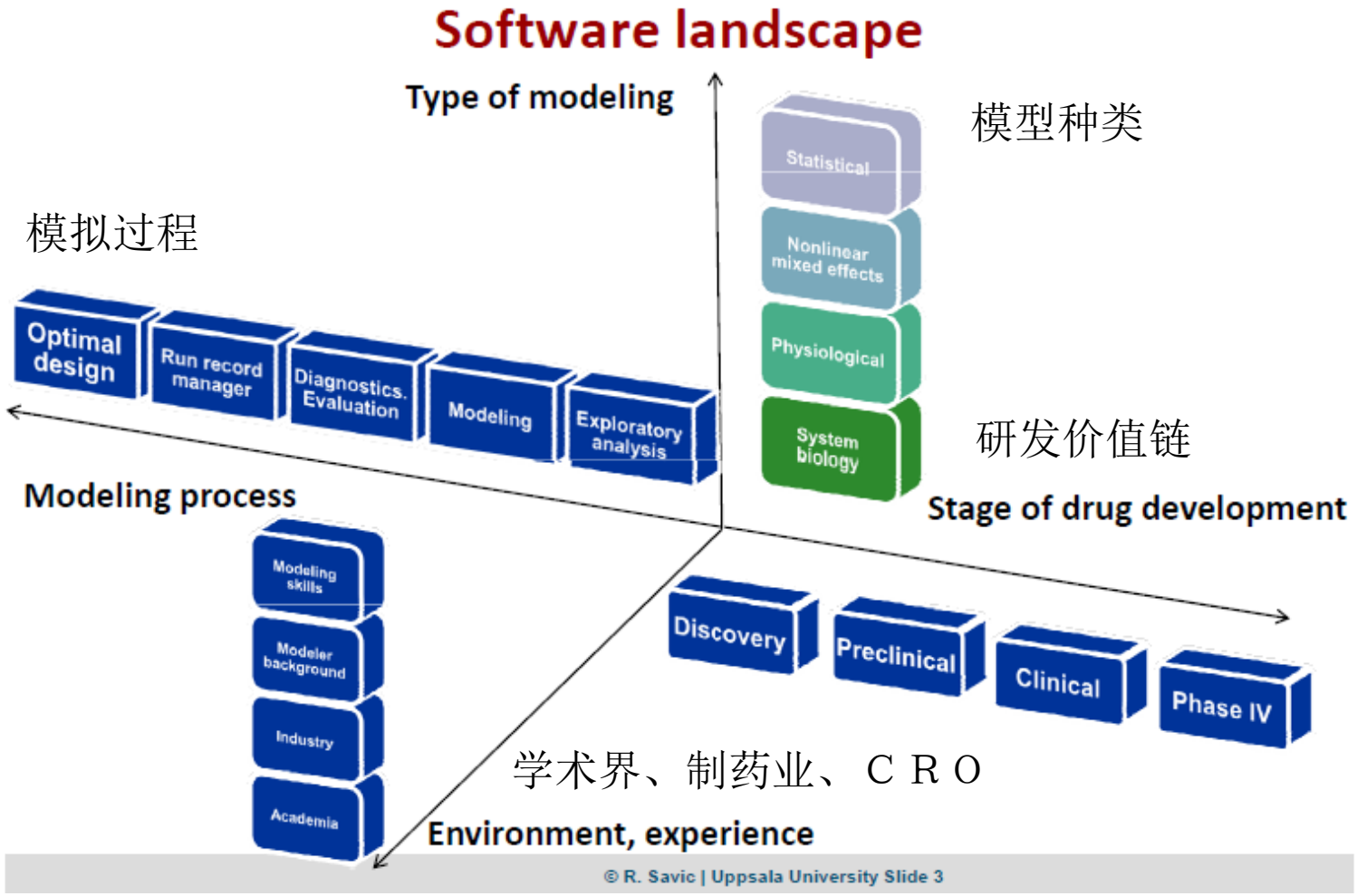
Modified from Piet van der Graaf

R Lalonde ISQP 2011

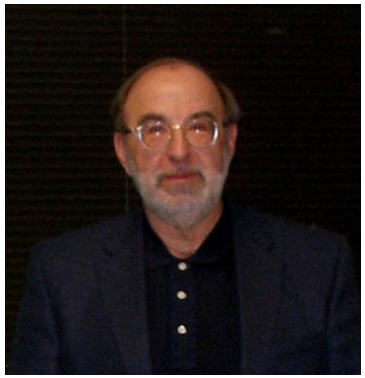


5

# 新药开发中软件的应用空间Space of Pharma R&D Software



# 效应模型的种类Types of Response Model



- 可逆性
  - 直接
    - 快速
    - 缓慢
  - 间接
    - 合成、分泌
    - 细胞迁移
    - 酶诱导
- 随时间变化
  - 耐受和超敏
  - 昼夜节律 (基线)
  - 年龄 (器官成熟/老化)
  - 疾病进程
- 连续性变量

- 非可逆
  - 化疗
    - 肿瘤细胞的杀灭
  - 酶的失活
    - 基于机制的抑制
    - 抗血小板凝集药
    - 质子泵抑制剂
  - 一些药物毒性
- 时间恒定
- 非连续性变量
  - 二分类 (e.g. yes/no)
  - 有序 (e.g. 疼痛分级评分)
  - 计数 (e.g. 癫痫发作次数/月)
  - 无序 (e.g. EEG 睡眠分期)
  - 时间事件 (e.g.生存)

# 统计模型的进展

UCSF Prof Sheiner

确定 ⟶ 随机

$$C_{ij} = \hat{C}_{ij}(1 + \varepsilon_{pij}) + \varepsilon_{aij} \qquad P_i = \hat{P}\exp(\eta_{Pi})$$

个体 ⟶ 群体分析

$$TVP = \theta_n \cdot \prod_{l}^{m} \left(\frac{cov_{mi}}{ref_m}\right)^{\theta_{(m+n)}} \cdot \prod_{l}^{p} \theta_{(p+m+n)}$$

变异 ⟶ 不确定性

**Bayesian**

$$p(\theta|y) = \frac{p(\theta)\,p(y|\theta)}{p(y)} = \frac{p(\theta)\,p(y|\theta)}{\int p(\theta)\,p(y|\theta)\,d\theta}$$
$$\propto p(\theta)\,p(y|\theta)$$

Bayesian推断

- 应用以往知识综合新数据 (客观证据和主观判断)
- 定量地描述概率的不确定性

# 选择分析软件的考量因素

## *从新药开发的产业视角*

- 是否能符合药物监管部门的要求？
- 是不是安全，能归档，追溯和接受审核？
- 是不是机器依赖性，操作系统依赖性？
- 价格是否可以承受？
- 速度快不快？
- 人员受训时间的长短？
- 能不能处理大批量数据？
- 有没有大批同道的支撑？
- 能不能随心所欲地作数据整合和科学作图？
- 统计程序包是否满足工作需要，能不能建模和模拟？
- 与其它软件的输入输出界面（如ＳＡＳ、ＬａＴｅＸ）是否广泛？
- 能不能编程、交互、可视？

# R: Regulatory Compliance and Validation Issues
# A Guidance Document for the Use of R in Regulated Clinical Trial Environments *August 17, 2008*

- Qualification and Validation of Systems for 21 CFR Part 11 Compliance
  - Part 11: Electronic Records, Electronic Signatures
  - Validation
- Software Development Life Cycle
  - Source Code Management; Testing and Validation; Release Cycles, Availability of Current and Historical Archive Versions; Maintenance, Support and Retirement; Qualified Personnel; Physical and Logical Security; Disaster Recovery
- CFR Part 11 Compliance Functionality
  - Accurate and complete copies of records for inspection, review, and copying
  - Limiting system access to authorized individuals
  - Use of audit trails, operational system checks, authority checks, device (e.g., terminal) checks
  - Controls for Open Systems

# Why R?

- The R project group has published the above document, describing their efforts to validate the R software and the development process.
- R is open-source, and all modifications from version to version are transparently defined to the user community. There's no guessing as to what has been modified when a new release is issued.
- R is much more widely used than S-PLUS, and has a very active discussion list. Any issues with the software are quickly identified and shared with the community
- It is necessary to ensure that installation and use are properly qualified. This process is defined within each company's own SOPs.

Marc Gastanguay

# R语言的普及程度和发展趋势

在新药开发中以**R**为工具不断出现，举几例

- R2WinBUGs (Metrum Institute)
- XPOSE (Uppsula University)
- QTc analysis (FDA pharmacometrics)

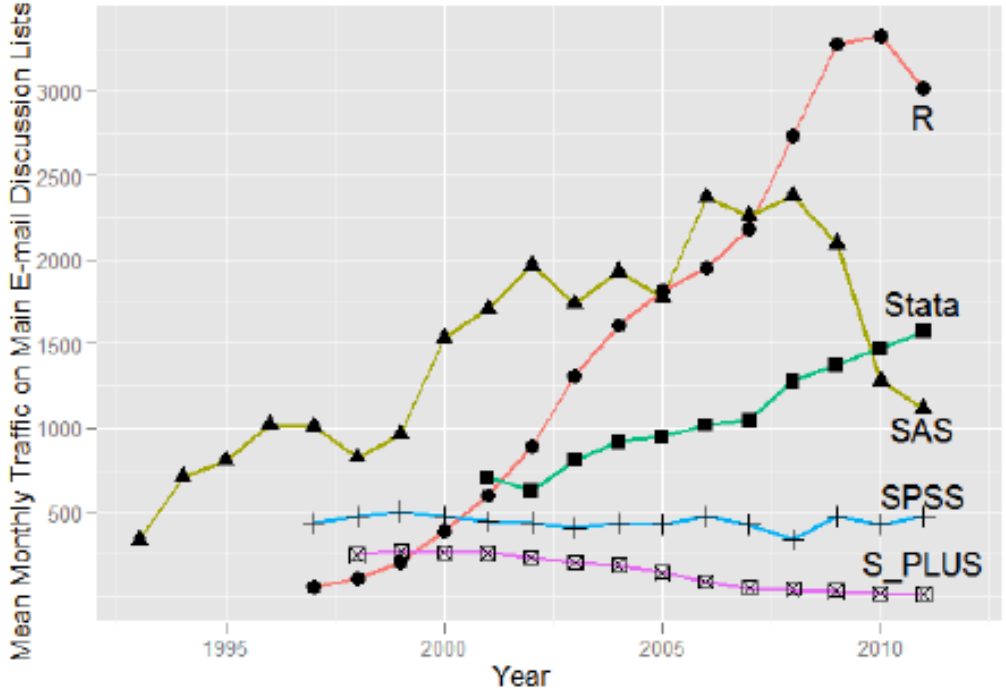
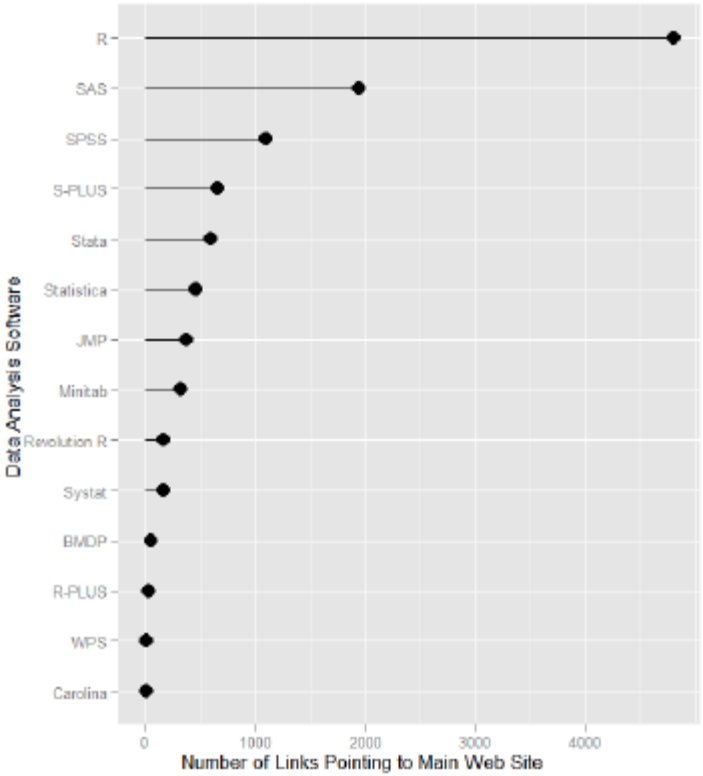
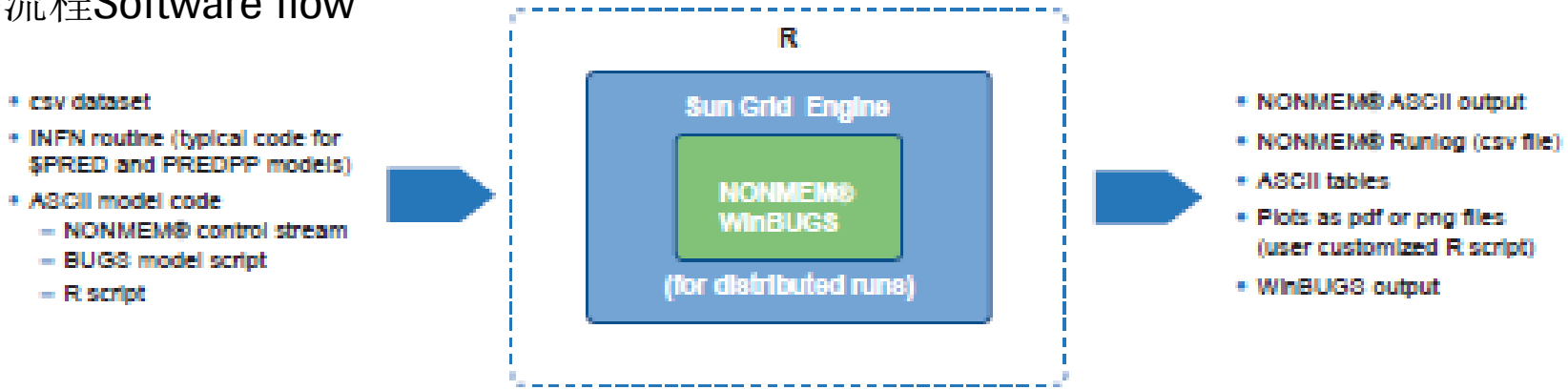



Figure 1. Plot of listserv discussion traffic by year (through 7/31/2011).

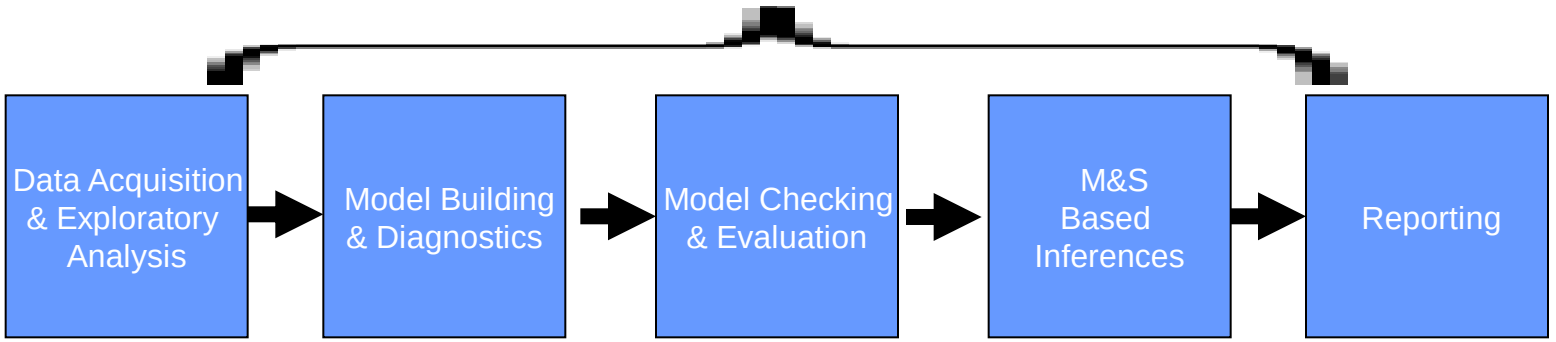


# R语言在新药开发数据分析系统的应用实例
## 定量药理师工作流程的MItools R 程序包

软件流程Software flow




工作流程Workflow

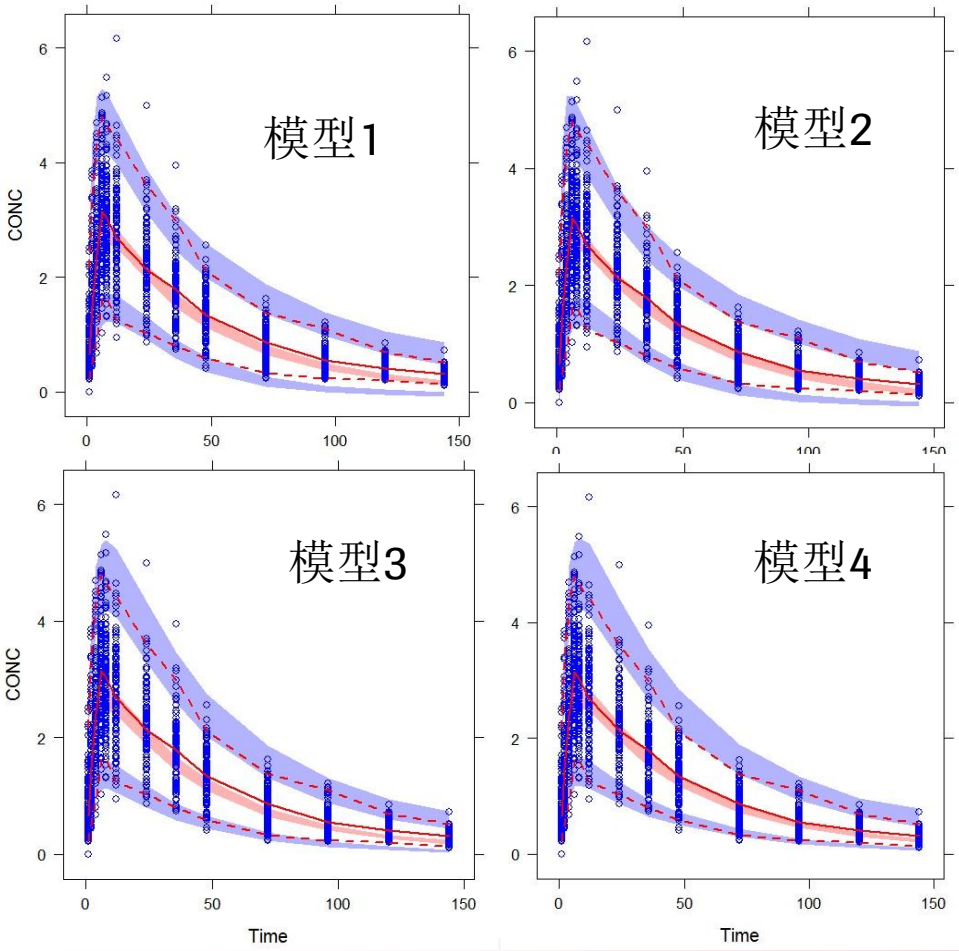



Key points: audit trail, version control (svn)

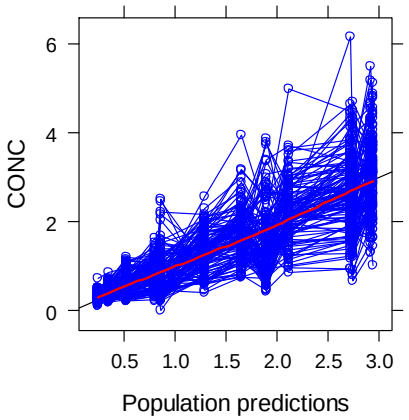
# R语言应用实例1

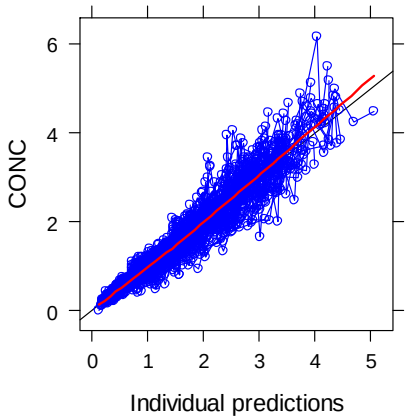
## 模型拟合的视觉预测检验（VPC）和诊断图

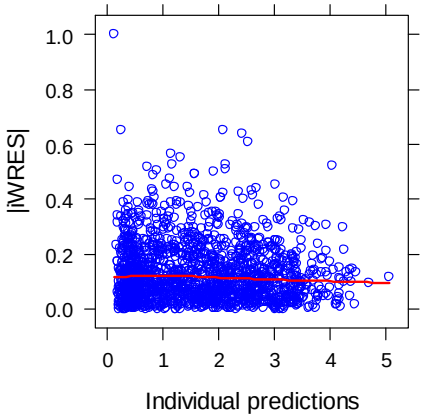
VPC

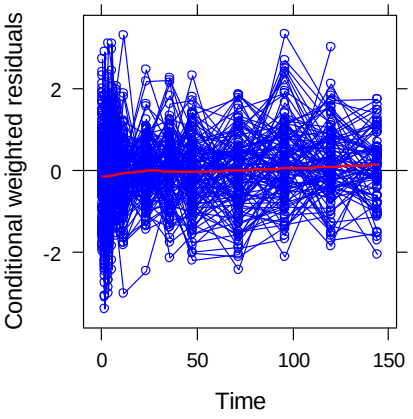



Diagnostic plots



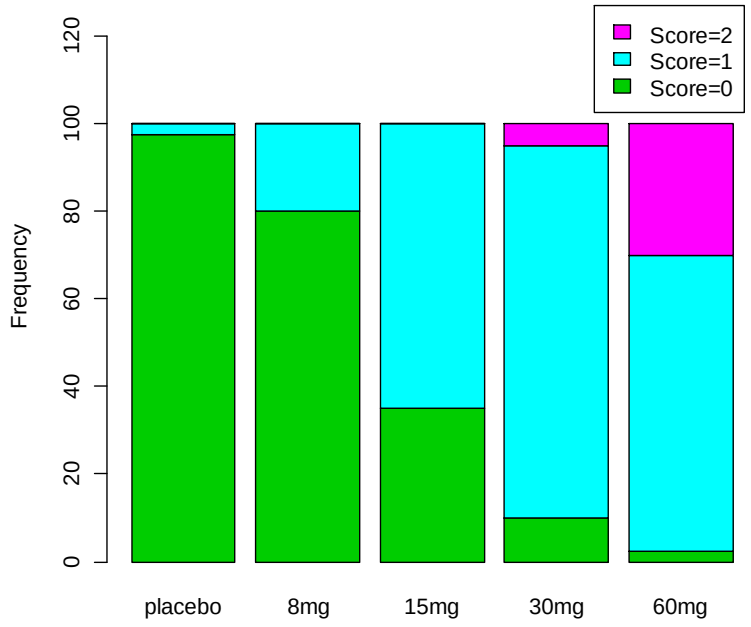



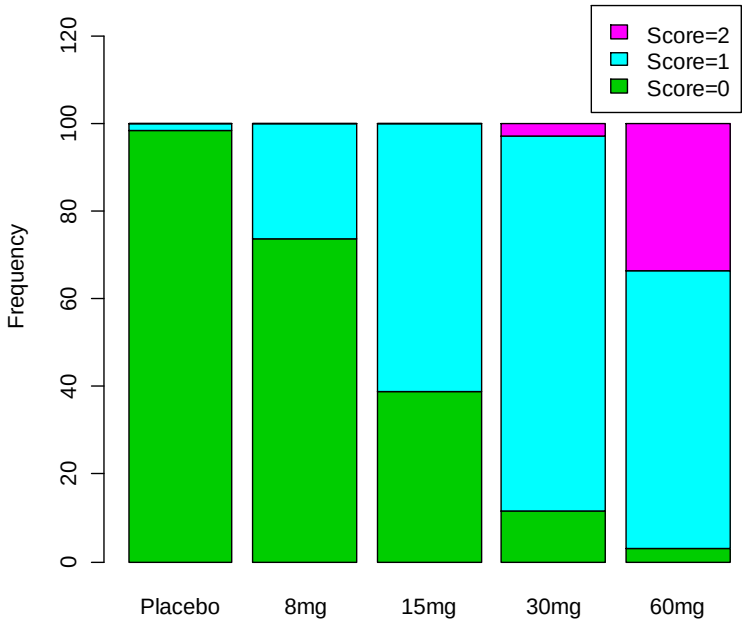

上海中医药大学王鲲博士提供

# R语言应用实例2 Categorical Analysis Model Qualification
## 有序数据效应模型的视觉预测验证

观测到的有序（等级）数据




模型预测的相应的数据



# R语言优点与缺点

- Fast and free.
- State of the art: Statistical researchers pro Advantages vide their methods as R packages. SPSS and SAS are years behind R!
- 2nd only to MATLAB for graphics
- Mx, WinBugs, and other programs use or will use R
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

- Not user friendly @ start – steep learning curve, minimal GUI.
- No commercial support; figuring out correct methods or how to use a function on your own can be frustrating.
- Easy to make mistakes and not know.
- Working with large datasets is limited by RAM
- Data prep & cleaning can be messier & more mistake prone in R vs. SPSS or SAS
- Some users complain about hostility on the R listserve

摘自网上信息

# Cutting Edge Science必须多学科的精诚合作

## 医学，药理学，统计学，计算机科学。。。

为有志于此道年轻学者提供一本参考书