医工结合之路：草木蔓发，春山可望

—— 访清华大学统计学研究中心助理教授俞声

◆ 本刊记者 杨璐 李彦

清华大学统计学研究中心助理教授俞声

清华大学统计学研究中心助理教授俞声，主要的研究方向是基于电子病历的数据分析，他与数据科学研究院的合作始于数据院和清华大学临床医学院合作搭建清华临床医学科研数据平台（以下简称：医学数据平台）。“我们自己去跟医院谈合作，不一定能引起人家的兴趣，医学数据平台的搭建促使我们能够更好的使用医疗数据做研究。”和长庚医院深度合作，俞声认为医学数据平台功不可没。

俞声有着丰富的国外医疗数据统计分析经验相较于国外成熟的医疗数据体系，国内医疗数据的收集、开放、处理都面临着诸多困难。“早期我主要研究美国的电子病历，诸如退伍军人系统这类最优质的数据我们都能拿到，但是国内相关的环境和规定还不完善，医院大都不敢提供数据，”他为我们分析道。“另外，中文病历的分析难度也比英文更大。美国有非常完备的术语库，可以用来辅助识别病历中的各种医学概念，将文字转化为变量。中文没有类似完整的术语库。所以一方面我们需要开发额外的术语自动识别技术，另一方面也希望国家加大医学信息基础设施建设的投入。”

除了医疗数据的使用和规范之外，国外的医疗数据研究团队也有相对成熟的模式。俞声介绍说，美国研究医学问题的团队会配备生物统计学家，医学背景的人提出研究问题，生物统计学家会帮助设计实验、建立模型、排除干扰变量，并对结果的解读提供统计学指导。“但是在国内，医生很少和统计学家合作科研。”俞声说，“所以数据院和临床医学院搭建的医学数据平台是非常有意义的。联合医学专家和数据科学领域专家，大家术业有专攻。像我个人是有统计和计算机交叉的知识背景，与医学专家合作，我们可以发现许多新问题，从新的角度去解决问题。”

通过医学数据平台，俞声团队在肝癌、脑卒中自由文本数据的信息提取、病历文本挖掘等方面都与医院展开了合作。医院提出和临床直接相关的医学问题并提供数据支持，俞声团队则提供技术支持并进行统计分析，得出和临床相关的结论。“为了数据安全，原始病历数据不能离开医院，所以实际上我们要处理什么数据，都要往医院跑，路上会比较辛苦。”提及与医院合作的过程，俞声有不少感慨：“数据院正在建立的医学数据平台，可以实现数据脱敏，并允许清华 IP 远程读取存储在医院服务器上的数据，远程处理之后传回结果。相当于数据还是留在医院，但是我们远程就可以实现处理和分析，就不用每次都奔波往返于医院和学校之间了。”

平台还在不断搭建和完善，国内医疗与大数据结合的工作还有许多困难需要克服。俞声相信，数据院和长庚医院的合作模式将会给全国带一个好头。在数据院的推动下，“医工结合”这条路已然草木蔓发，春山终可望！