智能AI+医疗数据

张　路
杭州美腾科技有限公司　浙江　杭州　310051

摘　要：随着人工智能技术的进步，其在医疗领域的应用变得越来越广泛。本文探讨了人工智能技术在医疗数据结构化处理中的应用，尤其是如何通过高质量的数据报告和产品，帮助医院处理脱敏后的病历数据，从而提高数据的使用效率和科研水平。通过分析目前的数据采集方式以及AI技术的应用，本文提出了一种全栈式AI数据采集方式，旨在改善传统方法中的效率低下和数据质量问题。

关键词：人工智能；医疗数据结构化；大数据；电子病历系统；数据脱敏

1 医疗数据的挑战

医疗数据，特别是文字类报告结构复杂、专业性强，将医疗行业的数据大规模转化为机器可识别分析的数据，即结构化，一直是行业的难题。据美国临床肿瘤学会（ASCO）统计，美国也仅有不到3%肿瘤患者的数据被结构化用于研究，剩下97%的数据都闲置在医院信息系统（HIS）里或者病历病案室中。作为一家专注于肿瘤大数据分析与应用的科技公司，通过承诺提供高质量的数据报告和产品，和诸多医院、科室合作，帮助他们处理脱敏后的病历数据，使电子病历信息转化为科研级数据以及临床决策支持，并研发人工智能工具，实现了肿瘤大数据一站式解决方案。

2 人工智能在医疗数据中的作用

人工智能有两要素：一是算法，二是数据。算法很多都是开源的，一个新算法出来，很快就能变成方便调用的模块。这也在一定程度上降低了参与者的门槛。对于医疗数据，它是均一化、标准化的有效数据，也是医疗AI的核心——这是因为医疗AI的大数据和其他行业的大数据有所不同，可能对于其他行业，大数据更多的是强调量、广泛性、泛化性，代表各个群体的特点，比如说做人脸识别，需要不同性别、不同年龄的数据，但是对于AI，更多是强调标准化和均一化，因为不同层级的医生诊疗水平不同，医疗行业中更多的强调名医与指南，强调顶级三甲医院的向下复制。

数据是一个行业性痛点。据说国外人工智能企业正排队要进入中国，因为中国庞大的数据资源。可是，我们缺乏高质量、有临床标注的数据。这也就解释了，为什么云上已经汇集一批医疗数据，但分析的成果却少之又少，数据还仅仅是堆积。人工智能创业企业需要先解决数据的标注痛点，这需要医疗专家的支持。实际上，专家在医院看病的经验和知识都已经在病历记录中沉淀下来了，如何把它挖掘出来？让它产生价值？“现在电子病历系统中有大量的非结构化、半结构化的病历文书，系统无法对其进行查询，这就要解决半结构化、非结构化数据如何利用的问题。主要问题有以下6点：

（1）医疗机构生产性系统差异性大。

（2）医院各系统数据采集粒度与接口标准不一致。

（3）医务人员填写数据随意性大。

（4）费时费力。

（5）接口标准定义不完善。

（6）医疗数据质量不统一

3 数据采集技术的进步

医疗大数据采集的三个关键环节是：多源异构数据融合、数据清洗转换、数据脱敏。目前，医疗数据大多散落在各个系统，碎片化、低质量、孤立分散、类型多样、标准不一，而优质的大数据采集手段可实现异构数据融合及数据的初步清洗（数据的前治理），为后续的大数据分析及应用奠定坚实的数据基础。开发合规前提下的数据标准化集成采集平台，可实现数据较高质量的存储及随时调用。数据采集方式可分为：

（1）传统的采集方式：

将HIS系统中的患者数据，以传统的手工或半手工方式为主，医生根据每个病种所需数据进行人工补录，即事后整理录入到Excel、Access等标准化模板里，生成标准数据文件，再用SPSS、SAS等分析软件对数据进行统计分析，这是临床医生做科研的传统套路；其中生成标准数据文件的过程（也就是数据“人工结构化”）。但由于科研的数据收集与临床诊疗工作不同步，往往滞后于临床诊疗，于事后对所需的科研数据进行记录，难免丢失了重要的过程性数据。并且在进行转录的过程中需要大量的人工核对，难免产生因人工失误造成的科研数据与诊疗数据不一致的情况。随着医院科研任务的增

加，对临床科研数据采集的准确性、完整性及采集效率要求越来越高，传统的数据采集方式已不能满足要求。

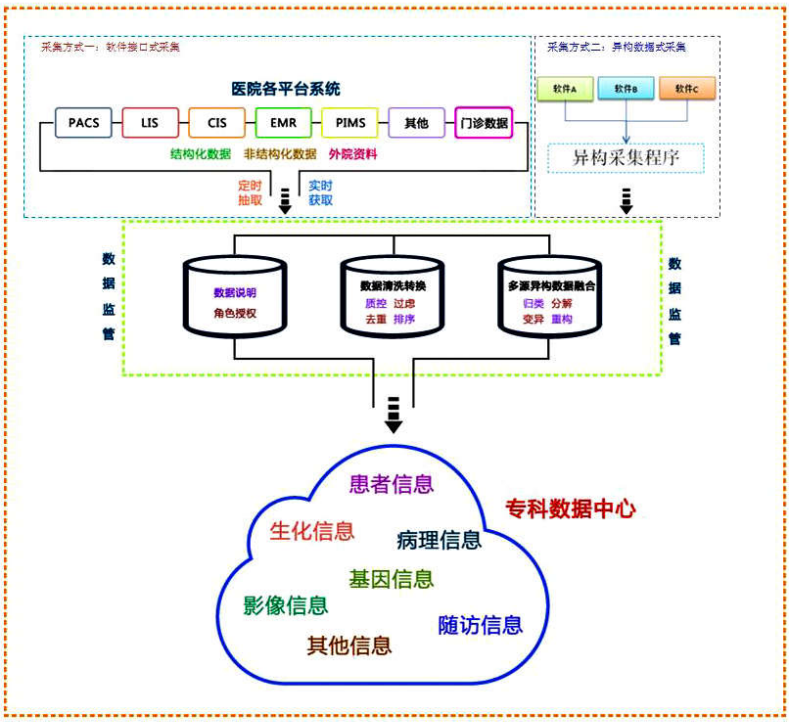
（2）人工智能采集方式：

在人工智能数据采集上，目前所有企业和公司都是接入医院数据平台，然后利用AI技术进行数据采集结构化，而这一过程缺少临床医生参与或暂用医生额外时间。

（3）全栈式AI数据采集方式：

在数据采集上大胆创新将数据采集融入到医生看病的过程中，既一边看病一看采集数据；数据采集直接在门诊采；帮助医生提高看病的质量与效率的同时又能采集数据；1+1＞2才是做医疗产品真正的初衷和目标。

全栈式AI数据采集与处理平台



安全保障体系

数据采集平台直接接入医院的HIS、EMR、LIS、RIS/PACS、超声、病理、外院资料等系统，减少中间环节。而医院内各个厂商所有业务系统的所有数据通过统一接口来集成是不现实的，这就需要一种方案，在不影响业务系统正常运转的条件下，在不需要任何业务系统进行接口改造的情况下，对数据进行采集，通过各系统底层数据接口和异构数据采集技术，把医院各业务系统数据实时抓取形成我们的数据中心(CDR)，再经过标准化、数据治理后，集成到大数据平台或数据中心里面。这一过程有效地隔离了对生产系统的影响，确保整个过程中不需要业务系统接口改造，数据实时秒级接入大数据平台。”

数据进入系统后，继续分工的思路，第一步是从医院抓取数据；第二步是转变成结构化数据对中文语义表达出来的术语进行依赖关系分析；第三步，基于医学信息学角度，以医学术语要求为依据，对医疗文本中的自然语言进行结构化处理，然后以关系型结构方式将这些语义结构存储到数据库中。结构化医疗文本主要特点在于对医疗文本中数据的层次结构关系进行规范。换句话说，就是尽可能的对医疗文本中的数据进行分解，以达到最小结构，并以此成为一个单元，使其在层级结构中都有相应的定位，从而能够进行结构化的录入和存储，并实现信息的快速查询与共享。从简单的部分开始，数据结构化的工作逐渐由“机器辅助人工”变成了“机器取代人工”。我国医疗术语缺乏标准化和医疗信息的复杂性，是机器取代人工的障碍。但前期大量人工检验的经验积累，让人工智能算法熟悉临床医生习惯的表述，明确了结构化基于的标准，形成了标准化的术语集，并开发了基于全栈式AI医学数据采集技术，包括自然语言（NLP）处理技术以及语音技术和“关键词”智能结构化系统。

4 六年探索捷径：从“海量数据”到“数据”再到“精准小数据”最终到达客观意义上的“大数据”

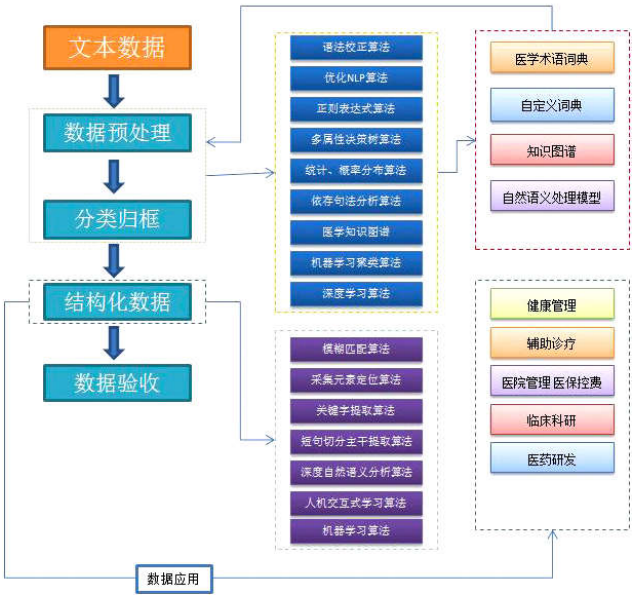
对于医疗行业，并不是数据越多、越广泛越好，也

不是数据遍布各个级别的医院更好，AI数据更多的是强调有精准的、标准化的、有效数据集合；而真正的医疗大数据是应该经过人工智能处理的医疗大数据，通过信息化互联互通以后，杂乱无章的数据仓库能通过人工智能的深度解析和结构化，把它变成一个数据宝库，可提取，可调阅、可分析。在数据库之上，再通过人工智能的医学建模和应用构建，产生注入临床辅助决策、影像辅助诊断、MDT个性化治疗等上层的AI应用果实。

借助计算机AI技术，可以从海量数据中快速找到有效信息并加以分析整理，将进一步提高科研效率。大数据应用探索是从“整合临床数据，根据科研人员的关注点展现信息”开始的。

最后总结为医疗数据采集4步法：第一步从医院系统（HIS、EMR、LIS、RIS/PACS、超声、病理、随访）海量数据中获取相关疾病信息完成“数据”的收集；经过深挖"数据"形成第三步一个个的“精准小数据”；随着精准小数据的累积最终终将形成客观意义上的“大数据”。

全栈AI数据采集流程与算法



文本数据结构化处理流程，适用于中文语言的文本数据结构化流程包括数据预处理、分类归框、结构化。“人工智能，有一大部分是通过机器学习完成了，给医疗数据处理带来了很大的帮助。”数据进入系统后，电子信息会自动解析、标准化再现并进行自动纠正校验；纸质信息会被扫描成图片格式然后由图片识别技术（OCR）识别成文本信息。在引入了深度学习技术后，计算机在复杂场景下也能快速适配。“比如错别字，标点符号等错误，系统就会自动更正。

目前我们95%的数据都能自动结构化，只剩下5%比较难的还需要人工。”人力劳动被解放，让数据处理能力大幅提升，一份病历的录入时间缩短到只需要5分钟，B超报告最快5秒，检验类报告更是瞬间完成。

临床数据中心平台是完全基于医院的内网服务器，合作医疗机构的原始数据保存在内网中，后续可将清洗、脱敏的数据传输到云上（基于阿里云的SaaS云服务），医生可以通过PC端或App进行访问。在合作医院和主任、医生在不断交流后，IT概念上的数据库和医疗行业的数据库有很大的区别。“我们IT行业的人说数据库，是说Sql Server、Oracle等，而医疗行业的人说数据库，是指从后台的存储到前台的可视化界面的整个解决方案。”

人工智能和处理后的医疗大数据结合，会产生许多新的帮助。可以为医院和科室的管理决策提供数据，也可以辅助医生的临床治疗。“患者来了，了解情况后，系统可以将过往类似患者的情况做一个归纳呈现给医生，辅助医生做诊断。

参考文献

[1]郭方栩.《计算机科学与技术的应用现状与未来趋势》电子技术与软件工程杂志,2018年第1期

[2]刘丹红,张林,杨喆,徐勇勇.《医学语言与临床数据标准化概述》中国卫生信息管理杂志,2014年第1期第四军医大学卫生信息研究所,白求恩国际和平医院核医学科

[3]李毅,保鹏飞,薛万国.等《中文电子病历的信息抽取研究》生物医学工程学杂志,2010年第27卷第4期

[4]王勇,李帅.《自然语言处理在医学文本挖掘中的应用》电子技术与软件工程,https://doc.taixueshu.com/journal/20191413dzjsyrjgc.html

[5]Center for Security and Emerging Technology (CSET), Small Data's Big AI Potential.https://cset.georgetown.edu/publication/small-datas-big-ai-potential