段永璇,甄天民,谷景亮,等. 大数据环境下肿瘤放疗的数据解析[J]. 中华医学图书情报杂志,2018,27(3):10-13.

DOI:10.3969/j.issn.1671-3982.2018.03.003

· 研究与探讨 ·

大数据环境下肿瘤放疗的数据解析

段永璇,甄天民,谷景亮,岳 媛,赵 悟

[摘要]介绍了肿瘤放疗数据的解析方法、肿瘤放疗数据的构成、肿瘤放疗数据解析的需求、肿瘤放疗数据实际的应用场景、肿瘤放疗数据解析意义、常用的数据解析方案,与解析工具和实际 TPS 系统数据解析的效果,总结了数据解析在肿瘤放疗大数据建设中的实际意义。

[关键词]肿瘤放疗;数据解析;大数据;医学信息;数据整合

[中图分类号]R730.55 [文献标志码]A [文章编号]1671-3982(2018)03-0010-04

Analysis of tumor radiotherapy data under the big data environment

DUAN Yong-xuan, ZHEN Tian-min, GU Jing-liang, YUE Yuan, ZHAO Wu

(Shandong Medical and Health Information Institute, Jinan 50062, Shandong Province, China)

[**Abstract**]Described in this paper are the analysis methods, compositions, demands, practical application environments, significance and analysis tools of tumor radiotherapy data. The TPS system data were analyzed, followed by a summary of the practical significance of data analysis in construction of tumor radiotherapy big data.

[**Key words**]Tumor radiotherapy; Data analysis; Big data; Medical information; Data integration

放射治疗是当今医学界治疗恶性肿瘤的三大主要手段之一。随着计算机数字化技术在医学领域的广泛应用,放疗设备及信息管理自动化程度不断提高,现代放射治疗技术正朝着精确定位、精确计划、精确治疗方向发展[1]。传统医疗服务模式更趋于数字化,数据已成为推动临床治疗及科研活动创新不可或缺的资料。自 2015 年起,美国国家卫生研究院、美国放射肿瘤学会、美国国家癌症研究所和美国医学物理学家协会召开的学术年会,都在持续关注放射肿瘤学在大数据时代的发展与机遇。国务院办公厅《关于促进和规范健康医疗大数据应用发展的指导意见》中提出推动健康医疗大数据资源共享开放,鼓励各类医疗卫生机构推进健康医疗大数据采集、存储,加强应用支撑和运维技术保障,打通数据资源共享通道。国内外的发展战略和研究目标,体现了各个层面对放疗大数据基础建设工作的重视,高效、合理地将各类数据进行有效整合,实现数据的高度集成,是大数据应用建立的关键,而数据解析则是实现数据整合的重要方法和手段。

1 放疗数据解析需求

在放射治疗过程中,数据主要通过 3 种途径产生:一是通过医院信息系统,如医院信息管理系统(HIS)、电子健康记录(HER)、个人健康记录(PHR)等产生的常规数据;二是通过放射治疗计划管理系统(Radiotherapy Treatment Planning System,TPS),如 Eclipse,Pinnacle,GammaPod 等计划系统产生的治疗计划数据;三是通过影像学检查设备,如磁共振成像设备(MRI,CT,PET)等产生、存储于影像归档和通信系统(PACS)中的影像数据。患者的信息数据除了包括性别、年龄、病症等常规信息之外,还包括放

[基金项目]山东省医药卫生科技发展计划项目"基于数据池的放疗异构数据整合架构体系的研究"(2015WS0175);山东省医学科学院面上项目"医学信息及数据服务云平台建设研究"(2016-01)

[作者单位]山东省医药卫生科技信息研究所,山东 济南 250062

[作者简介]段永璇(1979-),男,新疆石河子人,硕士,副研究员,研究方向:医学信息及数据挖掘,发表论文 20 余篇。

射影像、治疗计划、治疗方法、治疗规程、放射剂量、治疗方剂等非常规数据。数据形式多样，结构化、半结构化和非结构化数据同时存在，从而构成了大量的多源异构数据[2]。一方面，大量患者诊断治疗数据的产生，为临床医生的临床治疗以及科研工作的开展提供了有利的数据积累；另一方面，由于放疗临床数据的复杂性以及医院内部各种系统的多样性，加之各类厂商对自己治疗计划系统和设备的技术保护，不可避免地造成数据的多源性、异构性，势必会造成各类数据一定程度上的相互孤立，导致临床医生及研究人员难以全面地掌握和分析数据，对科研及临床治疗方法的进一步研究带来阻碍，也会造成临床数据分析的片面性[3]。

针对肿瘤放疗数据建设中的此类问题，数据解析工作就显得尤为重要。所谓数据解析，就是针对目标数据的结构特征，结合适当的分析方法，对数据进行详细研究并提取有效信息的过程。借助数据解析方法，可实现分散的临床数据整体化，实现不同系统、不同类型事实数据之间的快速转换及整合，有助于放疗临床试验数据的高效管理，以及临床科研及临床数据挖掘等活动的开展。

2　放疗数据结构

国际上通用的医学领域数字传输标准是 DICOM（Digital Imaging and Communications in Medicine）标准，在放疗领域的数字存储传输标准是 DICOM RT（Radiotherapy Objects）标准。作为 DICOM 标准的扩展，该标准定义了放疗领域的相关概念、流程和应用场景。DICOM RT 定义的信息对象主要包括 RT Image（放疗影像）、RT Dose（放疗剂量）、RT Structure Set（放疗结构集合）、RT Plan（放疗计划）和 RT Treatment Record（放疗治疗记录）5 部分[4]。DICOM RT 数据结构层次模型见图 1。

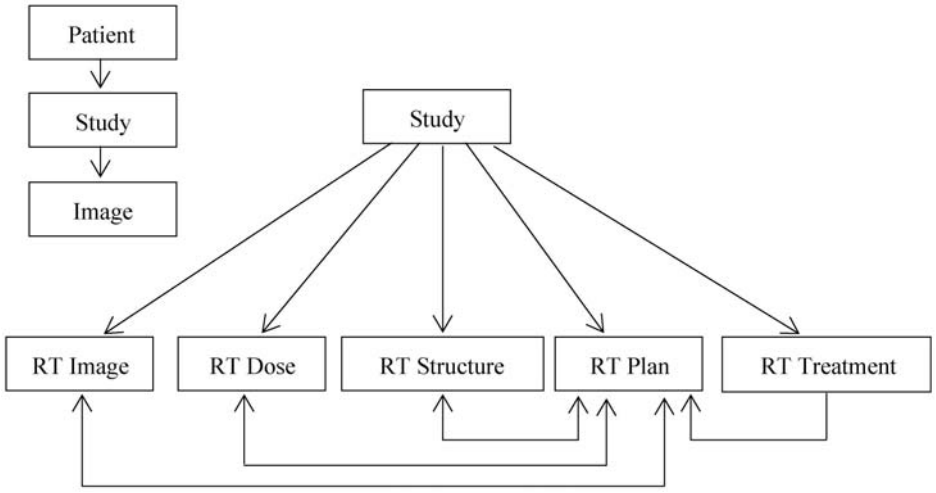

图 1　DICOM RT 数据结构层次

RT Image 是放疗图像以及图像相关的数据信息集合，包括 CT/MRI/PET 产生的图像以及数字重建图像、模拟机图像、射野图像等；RT Dose 主要用于传输治疗计划系统所计算的剂量数据集合，剂量的分布可以通过二维、三维的网格、等剂量线、剂量体积直方图（DVH）等表示；RT Structure Set 定义一个特殊区域的数据结构集，每个区域结构可以和一个或多个图像对象相联系，包括一些感兴趣区域（ROI、VOI）的定义、感兴趣点的选择（如剂量参考点）等[5]；RT Plan 是手工生成的计划报告、治疗计划系统及其他方式产生的计划报告数据集，包括外照射治疗、近距离治疗计划、分形、耐受性表、体位关系、控制点概念等；RT Treatment Record 是实际放疗过程中得到的记录数据集，包括记录信息的概要、所有治疗参数的记录、剂量计算、剂量测量记录等。

3　放疗数据解析过程

DICOM RT 数据是在实际的放射治疗过程中生成的。TPS 产生和涉及的数据信息量最为丰富全面，涵盖病人信息、图像信息、计划治疗信息等。因此进行 TPS 系统的数据解析，是解决肿瘤放疗数据

整合问题的有效途径。目前较为常见的数据解析方案是直接对 DICOM 文件进行操作，如利用 C 语言结合医学图像处理开源库（DCMTK）实现直接读取 DICOM 文件，获取相应的数据信息，或者通过 MATLAB编写代码对 DICOM 文件进行预处理，再结合 C 语言联合开发直接对 DICOM 文件进行数据操作，实现文件的分类。两者的共同点都是对 DICOM 文件进行数据操作，仅适用于对原始 DICOM 数据的读取和处理。现实情况是多数治疗计划数据是由各厂商提供的 TPS 系统产生，基于数据和技术保护的考虑，各厂商会采用自定义的数据封装格式将原始数据进行打包传输，而这种经过封装的数据大多是封闭的，难以直接应用，给数据的二次利用带来了很大的困难。以上两种方案均无法对 TPS 系统生成的数据文件进行直接处理，也无法实现 TPS 数据文件的传输、拆包、解析、存储入库等操作。针对此类数据处理的难点，笔者利用数据解析方法针对文件的结构特点设定解析规则[6]，采用 C++语言编写了 TPS 数据文件转换软件。该软件可对 TPS 数据进行底层处理，把封装的数据还原成原始数据，可用于二次解析的结构化数据，实现了 TPS 数据的自动化解析，解析后的数据存储在目标数据库中。图 2 是 TPS 数据解析后的部分数据包，包含了病人治疗计划的部分数据信息。

```
<PlanCreateDate>201410-05T13:09:13.1986491-04:00</PlanCreateDate>
<PlanCreator>st</PlanCreator>
<PlanModifyDate>2014-10-05T16:12:30.165-04:00</PlanModifyDate>
    <PlanModifier>st</PlanModifier>
    <PlanApproveDate />
    <PlanApprover />
  </Operations>
 <Stage ID="SelectPlan">
    <PatientInfo>
<UniqueID>10474e9b-c63b-4b7e-8373-0dd146646cd2</UniqueID>
      <PatientID>1047002</PatientID>
      <PatientName>zzz M J</PatientName>
      <PatientBirthday />
      <PatientSex>Female</PatientSex>
    </PatientInfo>
    <TreatmentInfo>
<TreatmentID>5025f559-9adc-4351-9ddf-294ee7a890c3</TreatmentID>
      <TreatmentName>1052013</TreatmentName>
        <TreatmentNotes></TreatmentNotes>
    </TreatmentInfo>
    <PlanInfo>
<PlanID>9f1cb660-4763-4024-b78c-cb875a5c04ae</PlanID>
      <PlanName>12</PlanName>
      <PlanDiagnosis></PlanDiagnosis>
<UniqueID>5025f559-9adc-4351-9ddf-294ee7a890c3</UniqueID>
      <Department></Department>
      <PatientID>GP_1047002</PatientID>
      <StudyDate>20121126</StudyDate>
      <StudyTime>151834</StudyTime>
      <StudyDescr>GAMMA POD</StudyDescr>
      <ReferPhysi />
<StudyModal>CT\RTSTRUCT</StudyModal>
      <AccessionN />
      <SeriesTime>110119.340000</SeriesTime>
      <SeriesDesc />
      <Modality>CT</Modality>
      <PatientPos>HFP</PatientPos>
      <ContrastBo />
      <Manufactur>Philips</Manufactur>
      <ModelName>Brilliance Big re</ModelName>
      <BodyPartEx />
      <ProtocolNa />
      <StationNam>HOST-7009</StationNam>
```

图 2 TPS 数据文档

在此数据集中可以看出，PlanCreateDate，PlanCreator，Stage ID 等为根节点；PatientInfo，TreatmentInfo，PlanInfo 则为 Stage ID 的子节点；PatientID，PatientName 等则为 PatientInfo 子节点中的具体数据。此类数据文件可采用 xml 数据解析方法，针对数据文档的树形结构，结合根、叶节点的特征编写解析遍历规则，即先扫描数据集的层结构，依次读取根节点信息，当遇到子节点后，继续扫描是否存在叶节

点，如果不存在则将子结点信息存储到当前的根节点下，如果存在则将叶节点信息存储在当前的子节点中，读取当前节点信息完成后，继续扫描下一个节点的内容，逐层获得数据集中的数据，并按对应关系进行存储[7]。通过以上方式，可获得 TPS 涉及治疗计划信息的完整的数据字段信息，包括患者信息、设备名称、DOSE 边界、放射剂量等。通过以上数据解析过程，可以得到 TPS 系统中的 VOI、DVH 等描述文件[8]、DICOM RT 的原始图像文件、Contour 数据文件等。DICOM RT 标准与 DICOM 标准[9]都采用 E. R 基本信息模型对实体进行抽象描述，使用信息对象定义的形式建立放射治疗数据模型，并用服务类的方式实现对放射治疗信息对象的操作。VOI、DVH 等数据描述文件，常常对应着大量浮点数据，不利于信息的检索和存档，因此需要通过数据的标准化建设[10]、数据降维等处理手段，构建相应的数据库及表。数据库设计过程中，由于 DICOM RT 标准中的图像是针对信息对象定义[11]，信息的存储或不同设备间的信息交换都是以 IOD 实例[12]来进行，所以数据库的设计尽量保持 IOD 的完整性，应体现 IOD 之间的关系[13]，可以按照患者、研究、系列和图像 4 个层次进行检索和管理，保持数据的完整性。因此，数据库的逻辑结构应与 DICOM 标准信息模型保持一致，易于体现各数据之间的联系。采用关系型数据库 MySQL 进行设计，尽量与 DICOM RT 标准保持对应，遵循统一的逻辑结构、信息对象关系、元素属性、属性值的表达方式等。对于原始的 DICOM 文件以及 TPS 系统生成的 DVH 文件，可采取扫描文件路径的形式，对文件名称及路径进行遍历，将文件的完整路径按照一定的逻辑结构对应地存储在数据库中。

4　结语

数据解析在当今大数据时代发挥着日益重要的作用。以肿瘤放疗数据电子化、高效化管理为目标，将数据解析应用于肿瘤放疗大数据建设中，可有效解决多系统肿瘤放疗数据集成化管理的难题。将医学信息分析理论方法与软件工程思维相结合，利用计算机编程和数据库技术，设计开发数据解析软件，符合医学大数据建设的发展趋势，为诊断、影像、治疗等多类医学数据资源的高效集成提供了一定的解决思路，对加快肿瘤放疗大数据的建设进程以及开展深层次临床数据挖掘等起到了积极的促进作用[14-15]。

【参考文献】

[1] 段永璇，邹晓艳，段秀梅，等. 数据解析在肿瘤放射治疗中的应用[J]. 国际放射医学核医学杂志，2015，39(6)：505-508.

[2] 詹国华，何炎雯，李志华. 智能健康管理多源异构数据融合体系与方法[J]. 计算机应用与软件，2012，29(9)：37-40.

[3] 刘　辉. 多源数据聚合系统及相关技术[J]. 电子技术与软件工程，2017 (19)：144.

[4] 雷　力，周凌宏，夏德国. DICOM-RT 标准解决放射治疗信息系统标准化通讯模型的构建[J]. 中国组织工程研究，2012，16(17)：3178-3182.

[5] 蒲立新，曲建明. DICOM 放射治疗中结构集文件的创建[J]. 中国数字医学，2011，6(2)：102-105.

[6] 段永璇，常文华，谷景亮，等. 医学信息采集的策略与方法[J]. 中华医学图书情报杂志，2016，25(9)：18-21，42.

[7] Yu C, Yao ZH. XML-Based DICOM data format[J]. Journal of digital imaging, 2010, 23(2): 192-202.

[8] 刘　芳，曹瑞芬，裴　曦，等. DICOM-RT 解析及在精确放射治疗计划系统中的应用[J]. 中国医学物理学杂志，2012，29(2)：3263-3266.

[9] 於文雪，张　惠，罗立民. DICOM 标准在放射治疗中的应用[J]. 中国医疗器械杂志，2002 (5)：352-355.

[10] NEMA DICOM PS 3.5 2015c. Digital Imaging and Communications in Medicine (DICOM), Data Structures and Encoding[S]. 2015.

[11] Bloch C. DICOM-RT Data Transfer of Structure Sets Between SRS Treatment Planning Systems[J]. Medical Physics, 2011, 38(6): 3637.

[12] NEMA DICOM PS 3.3 2009, DICOM-Information Object Definitions Part 3: Information Object Definitions[S]. 2009.

[13] NEMA DICOM PS 3.8 2015c, Digital Imaging and Communications in Medicine (DICOM), Network Communication Support for Message Exchange[S]. 2015.

[14] 马龙晖，刘海红，王　磊. 基于常用格式数据解析软件设计[J]. 电子世界，2017(11)：139.

[15] 左　翔，刘　方，胡学钢. 医学数据挖掘的探究与应用[J]. 中国农村卫生事业管理，2011，31(3)：268-270.

[收稿日期：2018-01-16]

[本文编辑：施沅坤]