肿瘤流行病数据可视化系统设计与应用

周奕洋 张 泽 钱 庆

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 介绍数据可视化研究的现状，常用可视化工具，现有数据可视化应用平台；在现状分析的基础上，设计出具有良好的交互性、动态性、可扩展性的肿瘤流行病数据可视化系统，可从多角度分析数据，使用不同可视化维度展示数据，为实际建设可视化系统提供理论基础和设计理念。

〔关键词〕 肿瘤流行病；数据可视化；数据分析；动态效果

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.05.011

Design and Application of the Data Visualization System for Tumer Epidemiology *ZHOU Yi－yang, ZHANG Ze, QIAN Qing, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper introduces the status quo of researches on data visualization, common visualization tools, and existing application platforms of data visualization. Based on the status analysis, it designs a data visualization system for tumer epidemiology with good interactivity, dynamics and extendibility. This system can analyze data from multiple angles, use different visualization dimensions to display data, and provide a theoretical basis and design idea for practical construction of visualization systems.

〔**Keywords**〕 Tumer epidemiology; Data visualization; Data analysis; Dynamic effect

1 引言

恶性肿瘤是威胁人类健康的最严重疾病之一，全世界和我国恶性肿瘤的发病率及病死率自 20 世纪 70 年代以来均呈现上升趋势，因而成为广泛关注的焦点。在流行病学的基础上研究恶性肿瘤在人群中的分布及其影响因素，以及探索恶性肿瘤的病因，制定相应的防治策略和措施并加以评价，最终达到降低人群恶性肿瘤的发病率和死亡率的目的一直是恶性肿瘤流行病学（Cancer Epidemiology）的主要研究内容[1]。在大数据时代，由于医学数据存在数量大、分析难的特性，需要强大的新技术用以提取各类有用的信息[2]，恶性肿瘤的相关研究也有了新的研究方向。可视化（Visualization）是利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。通过数据可视化技术可以将大量的数据集构成数据图像，同时把数据的各个属性值以多维数据的形式表示，从不同的维度观察数据，从而对数据进行更深入的观察和分析[3]。伴随着大数据时代的来临，进行数据可视化应用开发的人群也迅速扩大，促进了更加智能的数据可视化工具的出现。

本文研究肿瘤发病数据和可视化技术，设计从

〔收稿日期〕 2015－12－22

〔作者简介〕 周奕洋，实习研究员；张泽，在读研究生；通讯作者：钱庆，研究员。

〔基金项目〕 北京协和医学院“协和青年基金”项目：基于 HTML5 的肿瘤流行病数据可视化系统构建研究（项目编号：3332015055）。

性别、年龄、发病人数、所占比例等基本维度对肿瘤发病数据进行可视化展示的系统。扩展分析维度，从用户搜索行为、文献发表趋势以及维度之间的层次交叉引申维度上对数据进行深入的可视化分析。为肿瘤流行病领域科研人员与医疗从业人员提供肿瘤流行病发病数据的可视化展示支持，方便进行数据解读并识别肿瘤的流行特征趋势，有助于开展下一步的干预研究和预防治疗；为公众提供直观清晰的了解肿瘤流行病趋势的途径，有助于肿瘤防治与相关知识的普及。

2 研究现状

2.1 常用可视化工具

2.1.1 PC 端可视化统计分析工具 （1）Excel，可以进行数据的处理、统计分析及辅助决策，方便快速地与数据关联并实现多种基础图表，是优秀的可视化工具。（2）R，用于统计计算和统计制图的开源免费软件，R 语言拥有强大的社区和组件库，但相对复杂，需要较长的时间学习实践。（3）Gephi，用于各种网络和复杂系统的交互可视化与探测开源工具，主要用于探索性数据分析、链接分析、社交网络分析和生物网络分析等，通过简单的定义节点和边的操作即可生成复杂的网络可视化图谱[4]。（4）Tableau，一款上手简单的商业智能工具软件，可快速生成图表、坐标图、仪表盘与报告。

2.1.2 Web 端可视化工具 （1）D3. js，JavaScript 可视化库，通过使用 HTML、SVG 和 CSS 展示数据，兼容主流浏览器并避免对特定框架的依赖，提供强大的可视化组件，以数据驱动的方式去操作 DOM[5]。（2）Echarts，由百度开发的 JavaScript 图表库，兼容大部分浏览器，底层依赖轻量级的 Canvas 类库 ZRender。支持折线图、柱状图、饼图、雷达图等，提供图例、时间轴等可交互组件，支持多图表、组件的联动和混搭展现[6]。（3）HighCharts，JavaScript 图表库，支持柱状图、饼图、散点图等不同类型的图表，可以实时地从服务器取得数据并实时刷新图表[7]。个人及非商业用途免费，商业用途需购买许可。

2.2 现有数据可视化应用平台

2.2.1 商业数据可视化 （1）阿里指数，阿里巴巴公司提供的电子商务平台市场动向的数据分析平台，以阿里巴巴的交易大数据为依托，用可视化形式给出商业数据的各项指标与趋势。（2）友盟统计分析平台，移动应用统计分析平台，为移动应用开发商统计和分析流量来源、内容使用、用户属性和行为数据，其网站下“友盟指数”提供不同维度的移动端分析数据可视化展示。

2.2.2 公共数据可视化 （1）Google Trends（谷歌趋势），通过分析 Google 搜索结果，给出某一搜索关键词在 Google 被搜索的频率和相关统计数据，每一关键词的趋势记录图形显示分为搜索量和新闻引用量两部分，并有详细的城市、国家/地区、语言柱状图显示。（2）百度指数，类似谷歌趋势，以百度的搜索数据为基础的网民行为数据分享平台。提供关键词搜索趋势、监测舆情动向、定位人群特征等统计分析功能。

2.2.3 医学数据可视化 （1）谷歌流感趋势，通过分析人们的搜索关键词来跟踪预测全球流感趋势，意在通过提前预测未知疾病的流行，为疫情控制争取时间[8]。（2）GE 医学影像创新大赛，2011 年 9 月份启动，意在挖掘能够提高早期乳腺癌诊断水平的创意想法。该项目收到了 7 大类 500 个提案，其中五位获胜者获得种子资金用于将创意付诸实施。

2.3 现状分析

数据可视化工具多种多样，不同工具的侧重点不相同，可视化效果也不同，在不同的应用场景下应根据数据和需求结果选用合适的可视化工具。随着大数据时代到来，数据可视化展示平台也越来越多，在现有数据可视化服务中，商业数据可视化、公共数据可视化、医学数据可视化各有优势与不足，总结，见表 1。医学数据可视化以国外网站为主，且相对缺少具有良好展示性、交互性、从多维度分析数据、方便动态更新的医学数据可视化系统。

表 1 数据可视化服务对比

	展示效果	更新性	数据粒度	数据范围	收费性
商业数据可视化	较好，交互良好，维度较多	实时数据	较细	局限性强	收费
公众数据可视化	一般，一定交互性	实时数据	较粗	范围广	免费
医学数据可视化	一般，交互性较弱	静态数据为主	具体详细	专业性强	免费

3 肿瘤流行病数据可视化系统设计

3.1 概述

3.1.1 设计目的 设计采用可视化形式展现肿瘤流行病发病数据及其关联数据，提供直观展示与可交互操作，辅助科研人员、医疗从业人员进行数据分析、发现最新发病趋势以展开进一步研究与防治，也为公众提供良好的了解肿瘤流行病相关知识与趋势的途径。

3.1.2 数据来源 肿瘤发病数据来源于国家人口与健康科学数据共享平台 - 肿瘤转化医学专题服务，包括 ICD-O-3（国际肿瘤学分类）、ICD10（国际疾病分类）、年龄性别、出生日期、发病日期。肿瘤分类信息采用 ICD-O-3 对照 ICD10 总结整理而成，同时通过肿瘤分类关联 MeSH、CMeSH 词表，制作中英文检索入口词数据。通过检索入口词，查询整理 Google 搜索趋势、PubMed 文献发表量和 SinoMed 文献发表量作为趋势对比数据。

3.2 选用技术

3.2.1 HTML5 与 SVG HTML5 是对传统 HTML 的升级，将传统 HTML 中一部分繁冗的特性进行简化，对不足的特性进行补充和加强，使得 HTML5 能适应不同设备、不同浏览器，多媒体展示效果更好。SVG 是一种基于 XML 的描述二维图形的语言，作为可扩展图像，其分辨率和大小不固定，可以在不同分辨率设备上缩放和显示，且能保证图像质量，不会出现锯齿边缘或图像失真等情况[9]。

3.2.2 D3.js 和 Echarts 本次设计系统以网站形式展示，因此选用 JavaScript 可视化库 D3. js 和 ECharts。ECharts 具有良好的交互性和可定制性，提供多种不同样式图表，使用方便灵活；D3. js 能够提供大量线性图和条形图之外的复杂图表样式，与 ECharts 之间互相补充。

3.3 总体框架设计

可视化系统以网站形式展现，设计自下向上分为支撑层、数据层、功能层、表现层。支撑层包括网络、硬件和相关服务器软件；数据层存储肿瘤发病数据、肿瘤分类信息、Google 搜索趋势等数据；功能层提供搜索功能，以及从整体概览和局部透视的角度进行数据分析；表现层从多个维度可视化且具有交互性地展示数据。总体框架，见图 1。

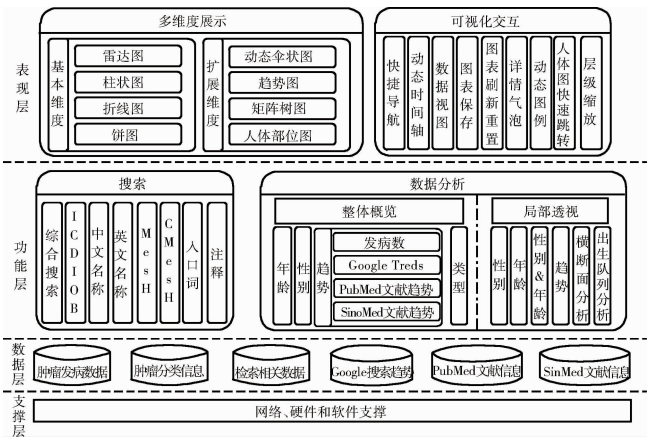

图 1 总体框架设计

3.4 功能设计

3.4.1 数据分析 从整体概览和局部透视两个方向对数据进行分析。整体概览对肿瘤数据进行全局分析，从年龄分布、性别对比、趋势对比、类型分布几个角度分析所有肿瘤的发病情况。局部透视针对每一类型肿瘤数据进行分析，从年龄分布、性别对比、性别与年龄分析、趋势对比、横断面分析和出生队列分析几个角度分析某一具体肿瘤的发病情况。具体内容：（1）年龄分布，分析高发年龄段和发病年龄在不同年份上的变化。（2）整体概览方向的性别对比，分析男性与女性整体发病率前 10 位

的肿瘤，对比男女性不同肿瘤发病率的高低，以及同一发病率排位上男性与女性不同的发病肿瘤。（3）局部透视方向的性别对比，分析男性与女性各自每一年份具体肿瘤发病率，对比不同年份肿瘤发病率的高低以及对比同一年份男性与女性不同的发病率。（4）趋势对比，发现和分析肿瘤流行病的发病趋势和特征，探索通过用户对肿瘤的检索关注（Google Trends）和学术研究文章发表趋势（PubMed 和 SinoMed 文献发表量）进行肿瘤发病趋势和特征探测，与流行病数据趋势进行比较，观察它们之间的关系。（5）类型分布，分析整体发病率前 10 位的肿瘤，得出高发肿瘤和其发病率对比。（6）性别与年龄分析，结合性别与年龄两个方面，将年龄以 5 岁间隔分段，比对不同性别下不同年龄段的肿瘤发病率。（7）横断面分析，分析同一时期不同年龄组或不同年代各年龄组的发病率、患病率或死亡率的变化，判断同年代出生的群体对致病因素暴露的时间和强度是否具有一定的相似性。（8）出生队列分析，将同一时期出生的一组人群作为出生队列，利用出生队列资料将肿瘤年龄分布和时间分布结合起来，用于探测肿瘤的年龄分布长期变化趋势。

3.4.2 多维度展示 针对不同的分析角度，从不同的维度展示数据，展示形式，见表 2。

表 2 展示维度

分析角度	展示形式
总体效果	动态伞状图
整体发病率前 10 位的肿瘤分布	雷达图（Radar）
	肿瘤发病人体位置图
整体前 10 位发病率与性别的对比	柱状图（Bar）
每年整体发病率前 9 位分布	饼图（Pie）
发病趋势	折线图（Line）
相关趋势对比	折线图（Line）
发病与年龄的关系	箱式图（Box Plot）
每年具体肿瘤发病率与性别的对比	柱状图（Bar）
具体发病性别与年龄关系	矩阵树图（TreeMap）
具体肿瘤发病横断面分析	折线图（Line）
具体肿瘤发病出生队列分析	折线图（Line）

3.4.3 搜索 考虑到受众群可能为科研工作者、医护工作者或公众用户，不同类型的用户搜索时使用的关键词也不同。所以整理和扩充已有数据，使得用户输入肿瘤的编号、中英文名称、俗称等都能够定位到需要查找的肿瘤。例如 ICD-O-3 编号为“C16”的肿瘤中文名称为“胃恶性肿瘤”，又称“胃肿瘤”或“胃癌”，要求用户输入“胃癌”时仍然能定位到“C16”。因此设计搜索功能提供不同的搜索范围，包括：综合搜索、ICD-O-3、中文名称、英文名称、MeSH、CMeSH、入口词、注释。当搜索的关键词只对应一种肿瘤时直接跳转至该肿瘤的详细信息页面；当搜索结果有多个时，以列表形式展现。

3.4.4 可视化交互 （1）快捷导航，当页面的内容较多、较长时需要一个快速定位的功能，因此设计在页面左侧用几个简单的图标表示对应项目，点击后能够直接跳转至指定位置。（2）动态时间轴，动态伞状图和饼图都设计使用时间轴，按照年份划分时间，数据随时间轴节点动态改变，可以暂停和选择指定时间点。（4）数据视图、图表保存、图表刷新重置，使用 ECharts 制作图可点击“数据视图”查看和复制可视化图表数据，点击“图表保存”保存选定的图表图像，点击“图表刷新”按钮重置图表。（5）详情气泡，鼠标悬停在图表上，显示对应位置的具体数据、时间、比例等详细信息。（6）动态图例，折线图、饼图提供动态图例功能，点击图例可以开关对应图块的显示，图表随图例开关动态刷新。（7）肿瘤发病人体位置图快速跳转，点击发病位置图中肿瘤示意图，可以直接跳转至该肿瘤的详细信息页面。（8）层级缩放，矩阵树图提供层级缩放功能，点击不同区块可缩放层级。

4 应用效果

制作了肿瘤流行病数据可视化网站（http://114.255.48.184/），效果截图，见图 2、图 3。

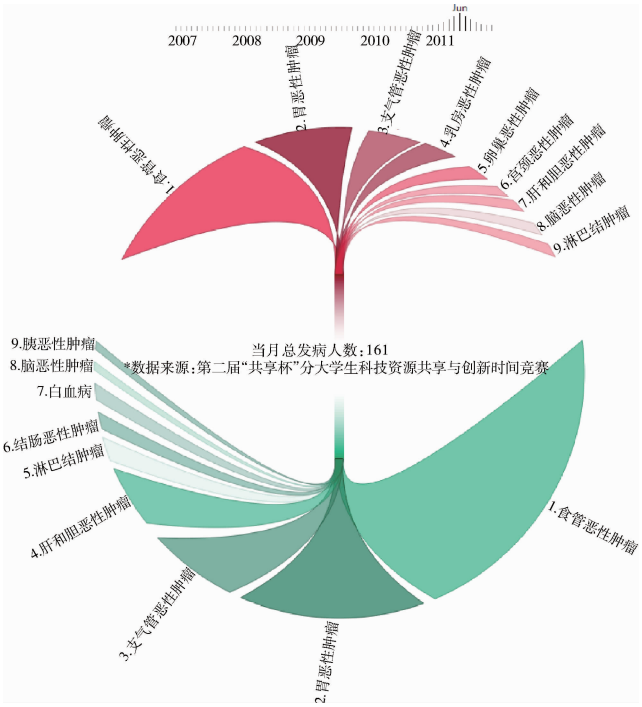

图2　动态伞状图

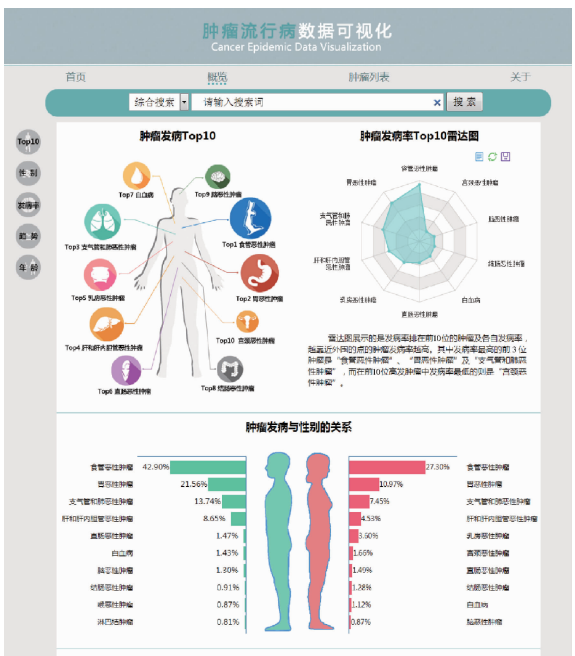

图3　快捷导航、发病人体部位图、雷达图、性别对比柱状图

5　结语

数据可视化是直观生动的数据展示方式，有助于数据分析、数据挖掘。本文中设计的肿瘤数据可视化系统具有良好的交互性、动态性、可扩展性，从多角度、多维度展示和分析数据。有助于肿瘤相关研究人员对数据进行分析，也有助于公众对肿瘤数据的认识和理解。在下一阶段的研究、设计和实践中将完善数据更新机制、扩展分析方法与维度、开发更多使用功能。并希望能以肿瘤数据可视化系统为基础，研究更多类型的医学数据可视化系统。

参考文献

1　詹思延主编. 流行病学［M］. 7 版. 北京：人民卫生出版社，2012.

2　罗志辉，吴民，赵逸青. 大数据在生物医学信息学中的应用［J］. 医学信息学杂志，2015，(5)：2－9.

3　杨昭. 多维数据可视化数据展示平台研究［D］. 济南：山东大学，2012.

4　彭琰，严莉. 基于 Gephi 的云南民族医药研究可视化分析［J］. 医学信息学杂志，2015，(2)：65－68，89.

5　Scott Murray 著. 数据可视化实战—使用 D3 设计交互式图表［M］. 李松峰译. 北京：人民邮电出版社，2013.

6　ECharts［EB/OL］.［2015－07－15］. http://echarts.baidu.com/doc/feature.html.

7　HighCharts［EB/OL］.［2015－07－15］. http://www.hcharts.cn/index.php.

8　俞国培，包小源，黄新霆，等. 医疗健康大数据的种类、性质及有关问题［J］. 医学信息学杂志，2014，(6)：9－12.

9　路莹，郝继英，陈锐，等. DRM 技术在机构知识库系统中的应用［J］. 医学信息学杂志，2014，(6)：27－29，48.

10　孟海滨，李立，王惠淑，等. 电子健康档案数据分析应用总体框架研究［J］. 医学信息学杂志，2014，(11)：2－7.