Journal of Data Science 18 (5), 875–888 DOI: 10.6339/JDS.202012_18(5).0004
December 2020

Quality Requirements for the Release of COVID-19 Data and Further Regulatory Suggestions

Yue Ding[1], Zhiyi Zeng*[1], Xichen Zhang[2], and Ao Chen[1]

[1]*School of Economics and Management, Southwest Jiaotong University, Chengdu, Sichuan, China*
[2]*Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, China*

Abstract

To surveil the development of COVID-19 is a complex and challenging issue. The foundation of such surveillance is timely and accurate epidemic data. Therefore, quality control for releasing COVID-19 data is very important, accounting for the releasing agent, the content to release, and the impact of the released data. We suggest that the quality requirements for the release of COVID-19 data be based on the global perspective that the goal of open epidemic data is to create a valuable ecological chain in which all stakeholders are involved. As such, the collection, aggregation, and release process of the COVID-19 data should meet not only the data quality standards of official statistics and health statistics, but also the characteristics of the epidemic statistics and the needs of pandemic prevention. The quality requirements should follow the unique characteristics of the epidemic and be scrutinized by the public. We integrate the perspectives of official statistics, health statistics, and open government data, proposing five quality dimensions for releasing COVID-19 data: accuracy, timeliness, systematicness, user-friendliness and security. Through case studies on the official websites of Chinese provincial health commission, we report the quality problems in the current data releasing process and suggest improvements.

Keywords *epidemic surveillance; epidemiological statistics; health statistics; official statistics*

*Corresponding author. Email: zhiyi0902@foxmail.com.

Received April, 2020; Accepted June, 2020

Journal of Data Science 18 (5), 875–888
2020 年 12 月
DOI: 10.6339/JDS.202012_18(5).0004

新冠疫情数据发布的质量要求及规范性建议

丁月[1]，曾智亿*[1]，张曦郴[2]，陈澳[1]

[1] 西南交通大学经济管理学院
[2] 香港中文大学工程学院

摘要

新型冠状病毒的疫情监控是一项复杂性强、挑战性大的工作。拥有及时、准确的高质量疫情数据是进行疫情监控的重要基础。新冠疫情数据发布的质量要求是个复杂的议题，需要考虑发布数据的主体、发布数据的内容和发布行为造成的影响。本文认为疫情数据发布的质量要求应该着眼于“以政府数据开放创造社会普遍参与的价值生态链”这一全局视角。基于此，新冠疫情的数据收集、汇总和发布过程应该遵循政府统计和卫生统计的数据质量规范、疫情统计数据的自身特点和防疫需要，形成适应自身特点的数据质量要求，并接受外界监督。我们综合了政府统计、卫生统计和政府数据开放的角度，提出新冠疫情数据发布的五个质量维度：准确性、及时性、系统性、友好性和安全性。通过对全国各省级卫健委官网的调查研究，我们发现了现有疫情数据发布过程中存在的质量问题，并提出了相应的改进建议。

关键词 卫生统计; 疫情统计数据发布；疫情监控；政府统计

1 引言

新冠疫情统计数据的及时、准确发布具有重要意义。对于国家而言，第一时间共享病毒基因组信息、搭建相关数据和科研成果共享平台、开展疫苗研发国际合作从而达到科学数据和信息共享，这是国际合作的前提和基础 (刘峣, 2020)。对于政府制定管理措施而言，及时、准确发布的疫情统计数据能够帮助政府充分掌握疫情发展态势、预测疫情走势，进而有针对性地制定管理办法、实现资源有效调度。对于个人进行有针对性的防护而言，通过清晰的数据，能够实时、准确地了解自己身边的疫情状况，在时刻提醒自己警钟长鸣的同时，也能避免不必要的恐慌 (The Lancet, 2020)。疫情统计数据的重要性在历史经验总结的过程中不断发酵，因此具有持续的影响力。比如分析 SARS 对铁路客运及货运运力的影响，为疫情下铁路运力的损失提供估计依据 (桂文林等, 2011)；又如基于埃博拉的疫情数据，多国之间建立了国际多边联防联控工作机制，进而构建信息报告机制，全面收集疫情相关数据，有效地抑制了埃博拉疫情的传播和流行 (贺祯, 2016)。鉴于新冠疫情数据的重要性，我们有必要确保疫情数据发布的质量。

尽管发布手段在不断提升，但是疫情数据发布的质量问题始终堪忧。以埃博拉疫情防疫为例，防疫过程存在的问题包括：关键数据字段缺失严重，数据传输不及时，没有统一的病患数据库，数据结构化程度低，不便于数据挖掘，数据更新不及时等 (Owada et al., 2016)。在本次疫情中，尽管国务院发布了《国家卫生健康委办公厅关于加强信息化支撑新型冠状病毒感染的肺炎疫情防控工作的通知》，明确要求“要深化‘互联网 +’政务服务”，依托各类在线政务服务平

*通讯作者。电子信箱：zhiyi0902@foxmail.com。

投稿于 2020 年 4 月; 接收于 2020 年 6 月

台，强化政务服务的一网通办 (国家卫生健康委办公厅, 2020)；但是在疫情发展前期，基层防疫人员仍然采用传统数据采集方式，如手工填报，经历各级卫健委层层上报，降低了数据回溯的效率，无法快速核验数据准确性，增加了数据管理成本 (中国信息通信研究院, 2020)。我们观察疫情数据发布过程还存在其他问题：各省份间确诊病例的数据维度不够丰富，地理空间数据开放不足，数据的可挖掘价值有待提高；不同层级、地区的卫健委发布时间、频次、形式不同，对数据提取和挖掘造成不必要的负担；不同地区间数据互通水平有待提高，对个人隐私保护水平亟待提升，如患者的姓名、照片，甚至身份证号码遭到不同程度的泄露。李月琳等 (2020) 对比了世界卫生组织（World Health Organization，WHO）美国疾病控制与预防中心（Centers for Disease Control and Prevention，CDC）和新加坡卫生部（Ministry of Health，MOH）在新冠疫情期间发布的数据, 发现数据质量不高的情况在国际范围普遍存在。这篇文章是相关主题的研究下国内目前仅有的一篇相关研究，论文从突发公共卫生事件角度阐述政府信息发布的特征，缺少从统计学视角展开的研究。本文聚焦“新冠疫情数据发布”这一概念，指的是包含了统计口径、统计范围、检疫手段等相关信息的新冠疫情一手数据，由政府相应主管部门通过官方网站、社交媒体、电视媒体等多渠道面向公众进行信息披露的特定环节。这是本文聚焦的核心，我们着重探讨这一环节的统计数据的质量要求。

从管辖权来看，新冠疫情数据是由中华人民共和国国家卫生健康委员会（以下简称“卫健委”）管辖、汇总并公布的部门级统计调查项目，报国家统计局备案 (全国人民代表大会常务委员会, 2009)，故应符合政府统计数据的规范要求。从内容来看，新冠疫情数据播报的内容专门针对中国范围内新型冠状病毒肺炎疫情，属于公共卫生安全的范畴，故应符合卫生统计数据的规范要求。从性质来看，政府及时准确地开放数据，预期可以带动公众参与数据利用，共同创造社会价值，从而实现政府数据开放全面生态系统的价值，故数据公布质量应符合政府数据开放的数据质量要求。然而以往文献只是单独地、割裂地从单一方面讨论数据质量的问题，而新冠疫情数据的发布过程存在具体的情境要求，其数据质量标准需要提炼，进而成为衡量政府卫生部门数据发布质量的重要依据。本文认为疫情数据发布的质量要求应该着眼于“以政府数据开放创造社会普遍参与的价值生态链”这一全局视角。基于此，新冠疫情的数据收集、汇总和发布过程应该遵循政府统计和卫生统计的数据质量规范、疫情统计数据的自身特点和防疫需要，形成适应自身特点的数据质量要求，并接受外界监督。

本文首先将从政府统计，卫生统计和政府数据开放这三方面梳理疫情数据发布的质量要求。其次，在此基础上，我们提出新冠疫情数据发布的质量维度，并针对现有卫健委及其下辖单位官方网站发布的数据质量进行调研。最后针对现有数据发布过程中存在的数据质量问题提出规范性建议。

2 疫情数据发布的质量要求

本节从政府统计，卫生统计和政府数据开放这三方面梳理文献，为制定新冠疫情数据发布的质量要求寻找理论依据。

2.1 政府统计数据的质量要求

在中国，依据统计调查管理的主体不同，政府组织的统计调查项目分成国家统计调查项目、部门统计调查项目和地方统计调查项目三类，这三类调查管理明确分工，互相衔接，互不重复。其中部门统计调查项目是指国务院有关部门的专业性统计调查项目，本文讨论的卫健委组织通报的新冠疫情调查就属于部门统计调查项目。在统计调查过程中，统计数据质量被喻为统计工作的“生命”，不仅关系到政府统计机构的形象和声誉，而且还会直接影响以其为依据所作的有关决策和结论的科学性与可靠性 (曾五一, 2010; 金勇进等, 2010)。特别的，中国自 2002 年 4 月正式加入国际货币基金组织（International Monetary Fund，IMF）的数据公布通用系统（General Data Dissemination System，GDDS）后，中国的统计数据采集、质量评估、公布等的标准都要与国际标准保持一致 (朱建平等, 2010)。

目前国际上还没有关于统计数据质量的统一定义。国际货币基金组织（IMF）提出的数据质量评估框架 (Data Quality Assessment Framework, DQAF) 集合了各国优秀的统计实践经验，参考了联合国官方统计准则和 GDDS 中的概念及要求，可操作性较强，因此被视为一个权威性的国际标准。DQAF 中确立了评估统计数据质量的五个方面：保证诚信，方法健全性、准确性、可靠性、适用性以及可获得性。朱建平等 (2010) 基于 DQAF 的理论框架，从全面质量管理的角度出发，将统计数据质量定义为：在数据收集、处理和公布等数据产生和公布过程中影响统计数据满足一般或特定用户需求的特征，其中包括数据收集过程中的客观性、适用性、准确性；数据处理过程中的方法健全性、可靠性、可比性；数据公布过程中的及时性、完整性和可获得性。金勇进等 (2010) 提出准确性是统计质量的核心，其次统计工作需要满足适用性、准确性、及时性、可比性、适用性、经济性、可得性和保密性。近些年，大数据成为可公开获取的新型数据源，大数据资源成为统计数据的有力补充 (李金昌, 2017)。因此国内外专家纷纷提出针对大数据的数据质量框架。广义数据质量框架下 (UNECE Big Data Quality Task Team, 2014) 数据质量维度的标准有：目的性、准确性、一致性与可比性、可得性和可释性。黄恒君 (2019) 基于广义数据质量框架下讨论了大数据在数据采集、数据处理与存储以及数据分析与推断的生产流程中可能存在的数据质量问题，并提出了将大数据融入传统调查的统计体系构建思路。

为了保证数据质量，统计数据的规范流程在不断完善，但是现有规范体系几乎完全针对国家经济运行领域，不能直接用于疫情数据的规范环节。1999 年国家统计局首次公布了国家统计局和省级统计局对主要统计指标数据进行质量评估的实施办法，针对国内生产总值等 12 项经济指标进行质量评估 (国家统计局, 2009)。数据的准确性也会经过缜密的流程审核得以保证，以国家统计局发布的国内生产总值 GDP（Gross Domestic Product）为例，其数据在部分之前会经过两次核算。比如，2018 年全年 GDP 数据的初步核算结果于 2019 年 1 月底公布，随后国家统计局根据年度数据资料、财政决算资料和部门行政记录等和全国经济普查结果修订 GDP 数据，于 2019 年 11 月底公布了最终核实的 GDP 数值。在大数据情境下，引入外部监督可以对各种统计数据生产过程进行监督，从而保证政府统计数据的质量。为了避免大数据中存在的数据误差在数据生成、采集利用过程中如滚雪球一般累计并被放大，大数据环境下的数据质量问题更不容忽视 (高敏雪, 2009)。同时，为了提升政府统计数据质量，应当从完善统计管理体制，统一与规范统计内容，加强统计队伍建设并增加工作经费，完善统计数据的质量诊断、评估和控制体系，推进统计建设并结合外部监督，在统计数据质量数据管理中引入数据质量成本，建立科学的政府官员

和统计工作绩效考核体系等多方面入手，多管齐下 (黄恒君, 2019; 陈建宝等, 2010)。

2.2 卫生统计的数据质量要求

从政府统计数据角度讨论质量要求时更多关注于统计数据的收集、处理及推断等环节的规范性，讨论数据规范重要性的情境更多基于经济背景。而卫生统计角度关注了卫生医疗情境，考虑了我国医疗数据采集、共享和访问等具体的问题。

完备的技术支持保证了卫生统计数据在现有医院管理中的作用。医院查看科室的运转以及医院的运行情况，做出管理决策都需要基于高质量的数据。为了保证数据的质量，卫生监督工作需要全面完善，具体包括建立统计术语的约定或标准，明确常用统计指标及相关计量单位，增大卫生监督检查的覆盖率，充分利用数据库和数据仓库工具等等措施 (张蓓蕾等, 2017)。公共卫生科学数据中心的数据共享更加应该注重数据共享的进度，数据采集流程的规范化，数据整合和共享环节应佐以管理条例和规范的约束，数据共享过程中关注用户使用的友好性、数据访问的可及性、可问性，同时注意数据安全和隐私的保护问题 (张英杰等, 2013)。

新冠疫情作为公共卫生突发事件，因此其数据发布环节对政府相关单位提出很高的要求。李月琳等 (2020) 针对新冠肺炎疫情下的政府发布信息进行了特征分析，文章聚焦在公共卫生事件中的政府数据开放的普适性上，认为疫情通报应该具有公开“通”晓，广而“报”告之意，但是对于疫情数据发布的统计需求没有深入研究。肖尤丹 (2020) 认为疫情信息发布主体在法律规定上存在冲突，新冠肺炎属于突发事件，突发事件的信息发布权由各级政府统一行使，但根据《传染病防治法》，传染病疫情发布权却仅属于国家和省级卫生行政部门，排除了政府及其他有关部门向社会公布传染病疫情的资格，因此如果将疫情数据发布简单地纳入政府统计或者卫生统计的范畴是会引起一定程度上的冲突。

2.3 政府数据开放的质量要求

政府统计和卫生统计的视角聚焦于数据本身，而忽略了数据发布以后产生社会价值的过程。在政府数据开放（Open Government Data，OGD）的议题下，数据质量同样重要，而且是关乎开放数据产生价值的关键。政府数据开放指的是在保护国家安全、个人隐私和商业机密的前提下，政府利用集成的网络平台，主动向公众提供无需特别授权、可被机器读取、能够再次开发利用的原始公共数据，以提升政府治理水平、促进经济发展、创造社会价值的公共服务活动 (郑磊, 2015)。OECD (2010) 结合政府数据开放的内容和固有特征，将政府数据开放的质量维度归纳为公共性、安全性、保证性、保障性、系统性、参与性、回应性、共享性和利用性。在开放数据的同时，元数据的质量也同样值得关注。

互联网时代的“+ 互联网”模式加速了政府数据开放的进程，也对政府数据开放提出了新的要求。“+ 互联网”模式是指以传统行业的既有业务为基础，主动利用互联网技术和理念，提高为用户服务的效率和质量的发展模式，更加强调从线下到线上的过程 (陈朝兵, 2019)。科技赋能，数据驱动，政府统计大数据应用中的“+ 互联网”模式可以延伸为“政府统计部门将原本存储在纸质文件或数据库中的数据主动上传至线上，供公众使用”的模式。因为政府掌握了国家大部分的核心数据，因此，虽然开放的政府数据不能带来短期经济利益，但是公开信息能够牵动多方的开发者参与到数据利用环节，以期实现社会层面的公共价值，这一过程旨在形成融入了政府、

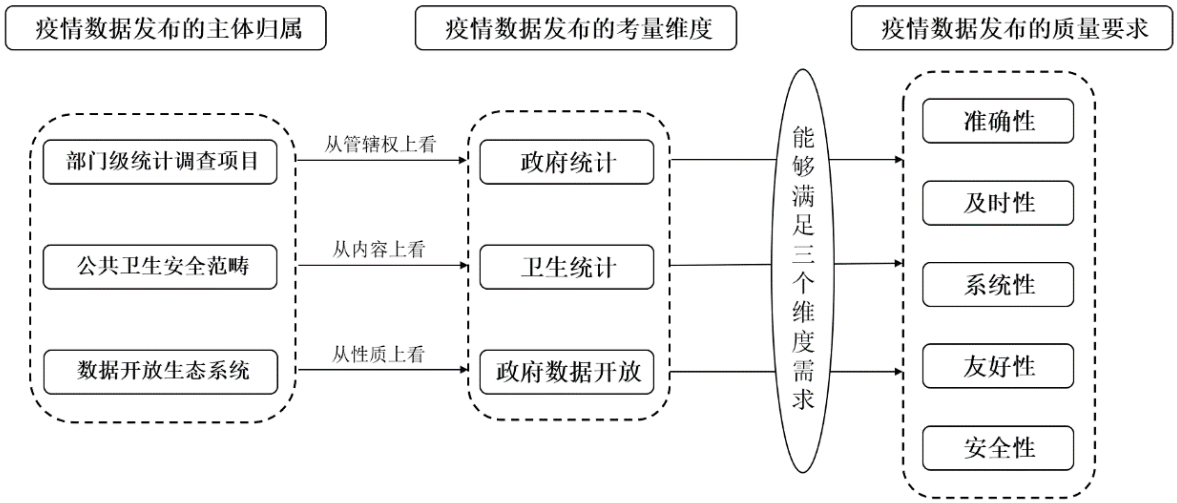


图 1: 新冠疫情数据发布的质量要求之概念维度。

开发者、社会多方参与的生态系统。在政府数据开放生态系统的价值产生过程，政府数据数量越多、质量越好，预期带动的参与者越多，创造社会价值的可能性更大。整个生态系统中注入的活力越大，良性循环的价值产生越可持续。考虑到价值链增值过程，免费开放的数据最终在社会层面产生的价值远远超过简单出售原始数据资源所获得的收益 (郑磊, 2015)。

生态系统的形成离不开传播链条的友好性，友好性主要体现在数据开放平台友好性和数据获取友好性。在数据开放平台的友好性上，主要由两部分构成，第一是易到达性，公众可以通过搜索引擎使用关键字搜索或者政府门户网站链接正确到达，不需要通过第三方网站从而减少误入钓鱼网站的风险；第二是操作便捷性，公众拥有基本的计算机操作技能便能够注册使用，避免公众注册和数据查询的程序过于繁琐。在数据获取的友好性上，主要由两个部分构成，第一是数据的可下载性，生态数据中的开发者可以直接下载数据而不需要利用爬虫等方法获取，将更多的工作重心转移至数据挖掘上；第二是数据的可机读性，例如数据格式为 XLSX、CSV 而非 PDF、PNG，不可机读性的数据即使能够下载也无法发挥应有的作用和效率。

综合上述文献整理，以往政府统计专注于经济领域的数据规范，应向其他领域延伸，新冠疫情造成的广泛社会影响使得问题延伸至卫生统计以外的范围，而政府数据开放概念下数据价值的产生过程值得我们借鉴。因此我们认为，单纯从政府统计、卫生统计或政府数据开放的角度都难以完整体现新冠疫情的数据发布质量要求，我们提出以新冠疫情为代表的卫生统计数据发布的质量标准应该着眼于政府数据开放的价值生态链，遵循政府统计和卫生统计的数据质量规范，形成适应自身特点的数据质量标准。综合以上三个方面的文献，我们提出疫情数据发布的数据质量要求。全文的概念框架如图1所示，并进一步提出疫情数据发布的质量要求。

3 疫情发布的数据质量要求

我们认为新冠疫情数据发布的质量维度应包含五个维度：准确性、及时性、系统性、友好性和安全性。从统计数据的全局角度而言，准确性和及时性是对数据的基本要求；从政府统计的角度而言，数据发布必须具有系统性和安全性，否则会对国家的公信力造成影响；从数据开放的角度而言，友好性是不可或缺的重要一环。

准确性

准确性是指在数据收集的过程中要保证一手数据的真实和准确，即疫情数据的统计值尽量贴近“真实水平”。做到准确性要求疫情指标的定义明确，疫情统计口径前后保持一致，从源头采集到一手数据。比如在疫情通报中准确区分感染者和疑似感染者，核对确诊人数是否包含无症状感染者，对输入型感染者准确划分归属地，这些问题需要从基层单位采集数据时就明确详细标准，并且在数据披露期间始终保持统计口径的一致。

及时性

及时性是指在第一时间发布和更新数据，尽力缩短数据搜集和发布之间的时间间隔。在疫情通报中要求各级卫健委至少做到每日疫情更新，部分高风险地区甚至可以结合自身需要做到每日两更。

系统性

系统性是指数据收集与开放的主体（国家卫健委及其下属子单位）应在各个数据要素（如统计指标，统计口径，检疫标准，数据披露周期，统计结果的披露格式等）和发布各环节（如数据收集，数据开放，数据质量监督与评估）保持系统性，一致性。

友好性

友好性是指用户获取和使用数据的便捷程度，例如数据可下载性和可机读性。在疫情统计数据开放的过程中，结构化的数据披露格式友好性较强，而部分数据以图片形式开放，明显不利于后续数据的加工与利用。

安全性

安全性涉及两个方面，第一是针对国家及个人安全，指数据不能对国家安全造成威胁或造成商业机密及个人隐私的泄露；在个别地区的疫情通报中出现了感染者的身份证信息及个人住址等信息，这是对个人隐私权的侵犯，违反了安全性的要求。第二是针对网站本身的安全，指应保障网站的运行平稳，提高网站安全等级，积极构建安全高效的信息平台。

4 卫健委疫情数据披露环节的代表性问题

基于我们提出的疫情数据发布的五个维度，我们进而跟踪现阶段卫健委及其下辖单位官网的数据披露情况，并进行调研。国家卫健委作为负责全国卫生健康工作的中枢机构，负责通报全国各省（直辖市、自治区）的疫情数据。整体而言，网站设计规范、数据披露比较详细。例如，网站不仅设立了疫情防控专区（见图2），而且披露内容全面规范。疫情新闻的标题格式统一，日期明确，让查询者可以迅速点击相应日期的疫情链接。在正文中，以 24 时为截止时间点，对上一日的确诊病例、重症病例、死亡病例、治愈病例、疑似病例、密切接触者、解除医学观察和接受医学观察这 8 个相关指标的人数进行了通报，内容简洁、明了且高效（见图3）。各级省市卫健委

图 2: 国家卫健委每日通报列表页。

图 3: 国家卫健委每日疫情通报详情截图。

网站的数据发布普遍做到了及时性要求，疫情发展阶段能够做到最低 24 小时的通报频率，部分省份在外部病例输入阶段有所放松，发布频率下降。然而很遗憾的是，各级省市卫健委网站数据发布的其他质量维度良莠不齐。虽然各省市体量不一，导致通报工作的复杂程度不同，但是在基础的信息披露规范上尚存在改进的空间。

具体而言，我们对除香港、澳门特别行政区和台湾省之外的 31 个省（自治区、直辖市）和新疆生产建设兵团的卫健委网站进行了详细调查比对，发现了一些问题。以省级卫健委官网为例，阐述情况如下。

4.1 统计口径不统一

在疫情初期，各级卫健委对于输入型感染者的归属地确认存在模糊空间，部分省市将其归为本地感染者，而部分省市则没有将输入型感染者纳入数据统计，造成“数据中空”的现象。此外，确诊人数是否包含无症状感染者，这一统计口径细节前后不一致，确诊标准前后存在略微差异。以湖北省为例，针对确诊人数是否应该包含无症状感染者存在反复定义，确诊的试剂指标和检疫标准中途变化。以上影响了数据的准确性，给数据复盘和数据质量监督带来额外的难度。

统计口径变化违反了疫情数据发布的准确性要求，使得发布数据的横向比较效力降低，基于这些指标诊断疫情进展并做出决策判断的可靠性降低。

4.2 数据最低通报区间不统一

相比国家卫健委披露的 0-24 小时新增数据的标准，部分卫健委网站统计区间不一致。部分省份存在跨日通报的情况，导致无法查阅 24 小时内新增疫情的数据。例如内蒙古卫健委通报数

据的起止时间不同，而且多次变化，疫情通报区间包含 7 时-次日 7 时，8 时-次日 7 时和 7 时-次日 8 时（见图4）。

疫情通报区间不一致，违背了疫情数据发布的系统性要求。导致了两个后果，第一是该省份无法进行有效地与自身进行纵向对比，第二是无法与其余省市进行横向对比，影响了跨区域汇总数据的精确性和权威性。

4.3 疫情通报缺乏规范醒目的疫情专区

以本次疫情为例，其传染性范围之广，后果之严重，备受全国（甚至全世界）的关注。而卫健委官网是政府机关、企事业单位、新闻媒体、以及普通大众关注疫情的重要网站。因此，非常有必要设立疫情专区，并将其设置于醒目之位置。但是，遗憾的是，一些省份卫健委官网并未设立疫情专区，而是将疫情播报夹杂在日常信息播报栏目中，比如“要闻动态”中，例如江苏省卫健委官网（见图5）。有些网站甚至还需要通过关键字查询才能查到，给相关各方了解疫情数据造成了较大的困难。

不同省份的卫健委网站未能设立统一格式的疫情专区，违背了疫情数据发布的系统性要求。这影响了公众及时、便捷地寻找到相关数据。

4.4 疫情发布格式不统一，非结构化数据不友好

以本次疫情为例，这是一个跨省、跨国的严重疫情。因此，应该调集所有能动员的研究力量，统一汇总全国各地的疫情数据，进行综合分析。但是，数据汇总的过程困难重重。其中一个很大的障碍便是：全国各级卫健委发布的疫情数据发布缺乏统一的通报格式。这对数据的采集和分析造成了很大的、不必要的障碍。

以详情页为例，存在图片截图（例如：山东省，见图6）、纯文字（例如：重庆市，见图7）、表格（例如：上海市，见图8）等多种形式。显然，从数据采集分析的角度看，图片展示的不可机读性不利于数据信息的提取，也不利于各类信息的汇总，从而影响了事后做各类分析的便利程度。而纯文字通报显得十分拥挤，缺乏结构化的呈现。

疫情发布格式不统一，使数据的格式化采集与分析产生困难，违背了数据发布的友好性要求。特别是在互联网时代，数据的传播依托于网络平台，不友好的数据意味着每一家网络平台在传播疫情新闻时需要增加额外的人力成本进行代码的修改和数据的校验，过程中也增加了引入传输噪音的可能性。同时，不友好的疫情数据使得外部科研人员需要投入更多的精力用于数据的抽取、转换和加载环节，带来更多的时间成本。

4.5 网络安全等级偏低，网站稳定性堪忧

全国各级卫健委官网披露的数据是官方数据，对社会影响巨大。因此，官网的稳定、畅通与安全非常重要。2020 年 2 月 4 日发生了印度黑客组织 APT 团体对中国医疗机构进行攻击的事件，这进一步提醒我们：卫健委网站应尽快提升安全等级，保障网站稳定畅通。这不仅关乎政府平台的运作安全，也关乎公众的信息安全。

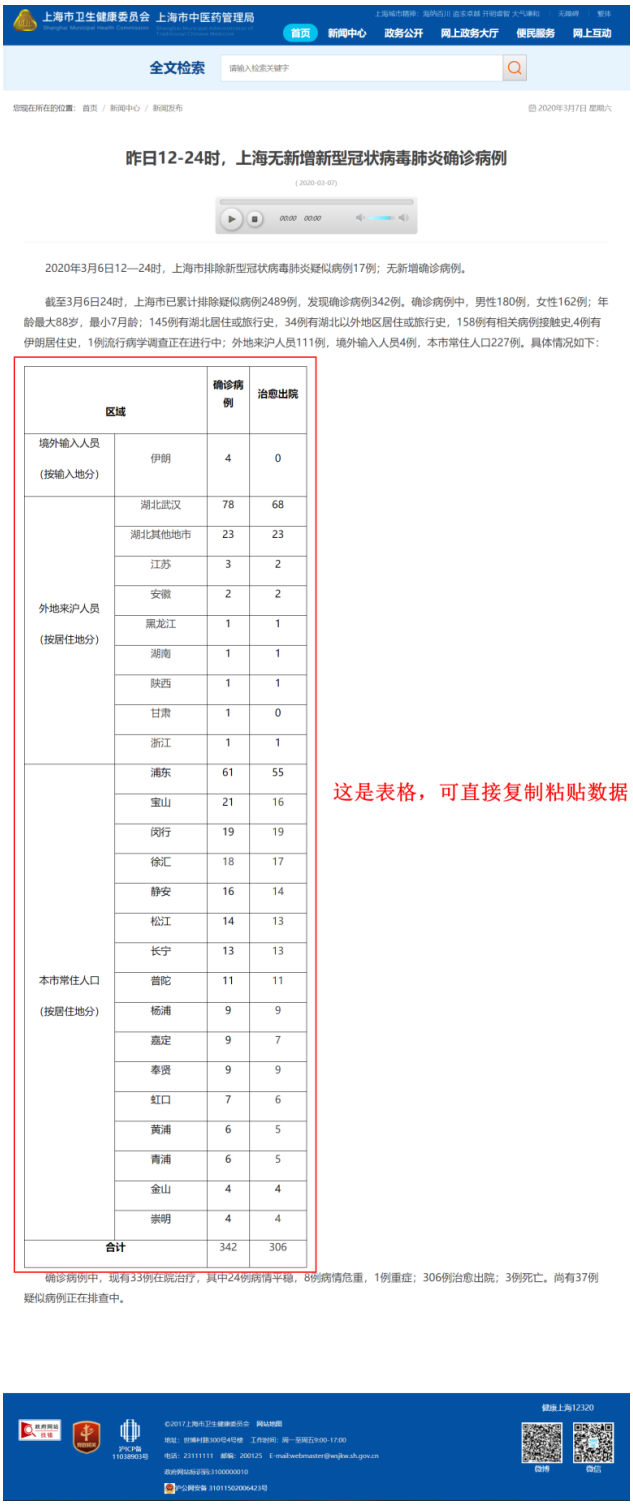
具体而言，提升网站安全等级的一个基本技术就是 https 传输，而不是 http 传输。前者更为先进，更为安全，而且也是非常成熟的技术。但是，我们发现一些地方卫健委官网采用了部分

图 4: 内蒙古卫健委每日通报列表页截图。

图 5: 江苏省卫健委官网首页。

图 6: 山东省卫健委疫情通报详情截图。

图 7: 重庆市卫健委疫情通报详情。



区域		确诊病例	治愈出院
境外输入人员（按输入地分）	伊朗	4	0
外地来沪人员（按居住地分）	湖北武汉	78	68
	湖北其他地市	23	23
	江苏	3	2
	安徽	2	2
	黑龙江	1	1
	湖南	1	1
	陕西	1	1
	甘肃	1	0
	浙江	1	1
本市常住人口（按居住地分）	浦东	61	55
	宝山	21	16
	闵行	19	19
	徐汇	18	17
	静安	16	14
	松江	14	13
	长宁	13	13
	普陀	11	11
	杨浦	9	9
	嘉定	9	7
	奉贤	9	9
	虹口	7	6
	黄浦	6	5
	青浦	6	5
	金山	4	4
	崇明	4	4
合计		342	306

图 8: 上海市卫健委疫情通报截图。

图 9: 浙江省卫健委官方网站。

图 10: https 传输下浙江省卫健委官网疫情专栏页面。

图 11: http 传输模式下浙江省卫健委官网疫情专栏页面。

http 的传输方式（例如：浙江省卫健委，见图9、10）。由于该官网部分采用了不安全的 http 传输，大多数浏览器（例如：360、Chrome 等）会进行拦截，并提示用户该网站不安全（见图11的左上角）。

网站非安全的传输使得访问的用户以及敏感网站数据都存在被黑客攻击的危险，不符合数据发布的安全性要求。作为政府发布的数据应当做到基本的传输安全和信息存储安全。

5 总结与建议

疫情监控是一件非常复杂、非常具有挑战性的工作。及时准确的疫情数据只是科学、快速决策的基本要求，鉴于新冠疫情数据的发布是政府作为发布主体，疫情监控是核心内容，发布数据

旨在引领社会多方参与，于是我们综合了政府统计、卫生统计或政府数据开放的角度，提出了新冠疫情数据发布的质量要求应扩展为五个维度：准确性、及时性、系统性、友好性和安全性。综合调研了全国各省级卫健委官网的数据发布质量之后，我们认为：虽然全国各级卫健委统计发布的疫情数据已经发挥了巨大的作用，但是尚存在巨大的改进空间。具体而言，我们建议全国各级卫健委：

1. 采用统一的数据收集、汇总和发布标准。具体的有：（1）统一统计口径：应及时修正因是否包含无症状感染者、疫情检测标准变化导致的统计口径变更，必要时完善多种统计口径数据并加强元数据的管理；（2）统一通报区间和发布频率：保持与国家卫健委一致的统计时间周期（0 时-24 时）和一致的发布频率（例如：每日）与时间（例如：早 8 点）。
2. 官方网站应设立疫情专区。各级官方网站的疫情专业应该统一位置，统一图文设计，统一各种公告的标题格式，甚至内容格式。
3. 疫情数据的发布采用统一格式。最好是各级官方网站都采用统一的可机读表格模板，甚至考虑数据批量下载的统一接口，这样利于整理、解读和分析。
4. 全面提高数据的安全等级。采取各种可使用之手段（例如：https 传输），保障网站安全稳定，保证政府平台数据发布的权威性和稳定性。

希望我国能够建设、拥有一套高效、全面、科学的疫情数据发布的质量体系，使得新冠疫情以及类似疫情能得到有效控制。

致谢

本研究项目收到泰康溢彩公共卫生及流行病防治专项基金，北大社科部研究基金，国家自然科学基金青年项目 (71702154，71702155 和 71802189) 以及四川省应用基础研究重大前沿项目 (2017JY0225) 支持。

参考文献

中国信息通信研究院, 2020. 疫情防控中的数据与智能应用研究报告. http://www.caict.ac.cn/kxyj/qwfb/ztbg/202003/P020200305495005485729.pdf.

全国人民代表大会常务委员会, 2009. 中华人民共和国统计法. 法律出版社.

刘峣, 2020. 全球战“疫”中国科技贡献“硬核”力量. 人民日报海外版：http://paper.people.com.cn/rmrbhwb/html/2020-04/02/content_1979694.htm.

国家卫生健康委办公厅, 2020. 国家卫生健康委办公厅关于加强信息化支撑新型冠状病毒感染的肺炎疫情防控工作的通知. http://www.gov.cn/zhengce/zhengceku/2020-02/05/content_5474692.htm.

国家统计局, 2009. 统计风采：新中国统计 60 周年图片. http://www.stats.gov.cn/ztjc/zthd/xzgcl60zn/xzg60ntjdsj/200909/t20090921_68900.htm.

张英杰, 王松旺, 杨洋, 苏雪梅, 2013. 依托公共卫生科学数据中心规范和强化数据共享服务. 医学信息学杂志, 34(5): 47-51.

张蓓蕾, 张帆, 匡圆, 秦洋洋, 马莎, 胡兵, 王晖, 张琪, 陈刚, 2017. 对卫生监督工作数据统计的思考. 中国卫生监督杂志, 24(1): 11-15.

曾五一, 2010. 国家统计数据质量研究的基本问题. 商业经济与管理, 2010(12): 72-76.

朱建平, 陈飞, 2010. 统计数据质量评价体系探讨. 商业经济与管理, 2010(12): 77-81.

李月琳, 王姗姗, 2020. 面向突发公共卫生事件的相关信息发布特征分析. 图书与情报, 2020(1): 27-33+50.

李金昌, 2017. 关于统计数据的几点认识. 统计研究, 34(11): 3-14.

桂文林, 韩兆洲, 2011. 季节调整本底线与 SARS 对我国铁路运量的影响. 铁道学报, 33(9): 10-18.

肖尤丹, 2020. 新冠肺炎疫情对公共卫生应急法治的重大挑战及对策建议. 中国科学院院刊, 35(3): 240-247.

贺祯, 2016. 联防联控创建发展与塞拉利昂埃博拉疫情防控多边救援分析. 解放军预防医学杂志, 34(1): 89-90.

郑磊, 2015. 开放政府数据的价值创造机理: 生态系统的视角. 电子政务, 2015(7): 2-7.

金勇进, 陶然, 2010. 中国统计数据质量理论研究与实践历程. 统计研究, 27(1): 62-67.

陈建宝, 陈谢斌, 2010. 政府统计数据质量问题及对策. 商业经济与管理, 2010(12): 87-91.

陈朝兵, 2019. 超越数据质量: 政府数据开放质量的几个理论问题研究. 情报杂志, 38(9): 185-191.

高敏雪, 2009. 从外部监督入手解决统计数据质量问题的努力. 统计研究, 26(8): 50-52.

黄恒君, 2019. 政府统计生产体系中的大数据融入探讨——基于数据源与数据质量的分析. 统计研究, 36(7): 3-12.

OECD, 2010. Reaping the benefits of cloud computing, web 2.0 and open data: OECD country experiences//Denmark: Efficient e-Government for Smarter Public Service Delivery. OECD Publishing.

Owada K, Eckmanns T, Kamara KBO, Olu OO, 2016. Epidemiological data management during an outbreak of ebola virus disease: Key issues and observations from sierra leone. Frontiers in Public Health, 4: 163.

The Lancet, 2020. Emerging understandings of 2019-nCoV. The Lancet, 395(10221): 311.

UNECE Big Data Quality Task Team, 2014. A suggested framework for the quality of big data. https://statswiki.unece.org/display/bigdata.