

# Structure and infrastructure of infectious agent research literature: SARS

Ronald N. Kostoff · Stephen A. Morse

Received: 23 February 2010 / Published online: 30 May 2010  
© Akadémiai Kiadó, Budapest, Hungary 2010

**Abstract** Text mining was used to extract technical intelligence from the open source global SARS research literature. A SARS-focused query was applied to the Science Citation Index (SCI) (SCI 2008) database for the period 1998–early 2008. The SARS research literature infrastructure (prolific authors, key journals/institutions/countries, most cited authors/journals/documents) was obtained using bibliometrics, and the SARS research literature technical structure (hierarchical taxonomy) was obtained using computational linguistics/document clustering.

**Keywords** Severe acute respiratory syndrome (SARS) · Coronavirus · Infectious diseases · Text mining · Bibliometrics · Citation analysis

## Introduction

The first pandemic of the twenty first century was the outbreak of Severe Acute Respiratory Syndrome (SARS) caused by the SARS-CoV coronavirus. As far as is known, this outbreak was not due to the deliberate release of the SARS-CoV coronavirus, but rather was a naturally occurring event. The appearance of SARS seems to have involved: (1) a zoonotic origin for SARS-CoV [e.g., horseshoe bats and/or Chiroptera as one wildlife reservoir (Janies et al. 2008)]; (2) transmission to intermediate hosts [e.g., civet cats, racoon dogs (Feng and Gao 2007)]; (3) human contact with these intermediate hosts in Southern China (Guangdong Province, November 2002) and subsequent cross-species

---

R. N. Kostoff (✉)  
The MITRE Corporation (Ret'd), 7515 Colshire Drive, McLean, VA 22102, USA  
e-mail: rkostoff@mitre.org; ronald.kostoff@pubpolicy.gatech.edu

S. A. Morse  
Bioterrorism Preparedness and Response Program, National Center for Infectious Diseases,  
Center for Disease Control, Atlanta, GA, USA

transmission of the coronavirus to humans (Feng and Gao 2007); (4) transmission of the virus through both non-hospital personal contact and hospital staff contact (Zhang 2007); and (5) global transmission of the virus via travelers from affected regions in Asia to other countries. SARS was eventually controlled through increased hygienic measures (e.g., face mask protection, frequent hand washing, living quarter disinfection), travel restrictions, and quarantine.

The implications of the emergence and spread of SARS are ominous. A new virus emerged that had not been identified previously. Due to a combination of several factors including: gestation period before the outbreak was identified and the causative virus discovered; bureaucratic delay in reporting the extent of the pandemic by Chinese officials; insufficient initial hygienic measures; inadequate hospital capacity for dealing with the full extent of treatment and confinement; and, insufficient initial travel and communication restrictions, SARS spread to about thirty countries before it was finally contained. During this outbreak, more than 8000 cases were reported to the World Health Organization (WHO), with a case fatality rate of approximately 10% (Hao et al. 2006).

How would these numbers change if SARS-CoV had been deliberately released by a bio-terrorist? No advanced technology would have been required. Infected intermediate hosts could have been distributed to multiple locations to serve as multiple sources for the infection. By the time the intentional nature of the outbreak was recognized, it would have spread globally, causing an overwhelming strain on healthcare resources throughout the world.

One of the less tangible weapons in the arsenal to combat both bio-terrorism and natural pandemics is intelligence. This has myriad forms, including the direct use of humans to access information, use of sophisticated computer systems to track infrastructure and resource movements, use of literature-related discovery techniques to help predict the development of engineered biological threat agents (Swanson et al. 2001) and to predict the emergence of naturally occurring pathogens, and development of technologies to detect, neutralize, shield and vaccinate against bio-warfare and naturally occurring agents.

To demonstrate the use of text mining to gain a better understanding of SARS and SARS-CoV, and how this information could be used to impact preparedness and response activities, a three-part examination of the global SARS literature was undertaken. In the first part, the global SARS research literature was retrieved, and the infrastructure of this literature (key researchers, institutions, journals, etc.) and medical structure (medical themes, inter-relationships among the themes) were identified. In the second part, the seminal documents of this literature will be identified, and summarized as a historical literature review (e.g., Kostoff et al. 2007). In the third part, literature-related discovery (Kostoff 2008; Kostoff et al. 2008) will be used to identify potential preventive measures and treatments for SARS.

## Background

### Text mining

Text mining is the extraction of useful information from large volumes of text. Its component capabilities of *computational linguistics* and *bibliometrics* were the main analytical techniques used in the present study.

*Computational linguistics identifies the main technical/medical themes of the database(s) being examined as well as the relationships among these themes*

Science and technology (S&T) computational linguistics (Hearst 1999; Losiewicz et al. 2000) is a process that underlies the extraction of useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from technical phrases that occur frequently. It also identifies relationships among these themes by grouping (clustering) the phrases (or their parent documents) on the basis of similarity.

Computational linguistics has been used for a number of purposes including: enhancing information retrieval and increasing awareness of the global technical literature (Kostoff et al. 1997; Greengrass 1997); discovery and innovation based on merging common linkages among very disparate literatures (Swanson 1986; Kostoff 2008; Kostoff et al. 2008); uncovering unexpected asymmetries in the technical literature (Goldman et al. 1999; Kostoff 2003); estimating global levels of effort in S&T sub-disciplines (Kostoff et al. 2000, 2004a); helping authors to increase their citation statistics by improving access to their published papers, which also may help journals increase their Impact Factors (Kostoff et al. 2004a, b); and tracking the impact of a specific research area across time and applications areas (Davidse and Van Raan 1997; Kostoff et al. 2001).

Evaluative bibliometrics (Narin 1976; Garfield 1985; Schubert et al. 1987) uses counts of publications, patents, citations and other potentially informative items to develop S&T performance indicators. Its validity is based on the premises that: (1) counts of patents and papers are a valid indicator of R&D activity in the subject area of those patents or papers; (2) the number of times those patents or papers are cited in subsequent patents or papers is a valid indicator of the importance or impact of the cited patent or paper; and (3) the citations from paper to paper, from patent to patent, and from patents to papers are an indicator of the intellectual linkages between the organizations that are producing the patents and papers, and knowledge linkage between their subject areas (Narin et al. 1994).

Evaluative bibliometrics can be used to: identify the infrastructure (authors, journals, institutions) of a technical domain; identify experts for innovation-enhancing technical workshops and review panels; develop site visitation strategies for assessment of prolific organizations globally; and identify impacts (literature citations) of individuals, research units, organizations, and countries.

A typical text mining study of the published literature involves the development of a query for comprehensive information retrieval, a search of the database using computational linguistics and bibliometrics, and an integration of the processed information.

## Materials and methods

### Database generation

The primary objective of this study was to identify the global research literature that was related directly to SARS. A secondary objective was to estimate the relative level of global effort in the sub-categories of SARS research, as reflected by the emphasis in the published literature. Because citation bibliometrics are an important tool, the SCI was selected as the database for most of the analyses. Additionally, only records classified as “articles” or “reviews” in the SCI were downloaded. Once the source database was selected (1998–early 2008), the iterative search approach of Simulated Nucleation (Kostoff et al. 1997) was used to generate the final search query.

The SARS research literature, as defined by the authors of this study, consists of published open-literature papers that focus on theoretical, laboratory, biological, clinical, and epidemiological aspects of SARS and its coronavirus, and emphasizes the original research literature accessed by the SCI. The SARS literature is defined operationally by the following query: [(SARS and (coronavirus or infect\* or virus\* or viral or epidemic\* or epidemiology or antibodies or antibody or vaccine\* or influenza or pandemic\* or outbreak\* or syndrome)) OR “sars patient\*” or “sars transmission” OR “SARS-CoV” OR “severe acute respiratory syndrome”].

### Document clustering

Based on recent text mining results (Kostoff et al. 2004a, 2006), document clustering was the only theme identification method used, and was performed using the Abstracts text only. Document clustering combined documents into groups based on their text similarity. Document clustering provided numbers of documents in each cluster, a proxy metric for level of emphasis in each taxonomy category. The approach used in this study was based on a partitioning clustering algorithm (Karypis 2004; Zhao and Karypis 2004) contained within a software package named CLUTO.

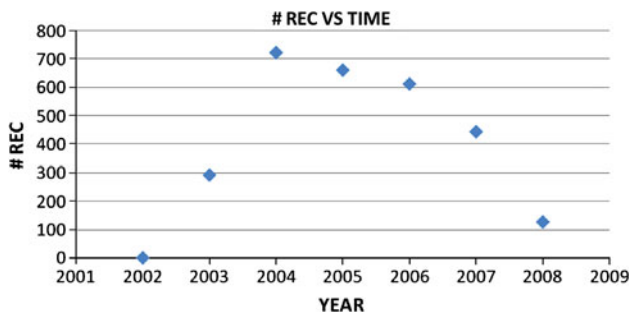
## Results

### SARS publications, 2002–2008

Figure 1 shows the number of SARS SCI articles (retrieved with the above query) as a function of time. Between 2003 and 2008, a total of 2874 papers were retrieved. For the major country producers of SARS research articles, the temporal production is as follows (country/number of SCI articles):

- 2003: China (159); USA (52); Canada (30); Singapore (23).
- 2005: China (273); USA (168); Taiwan (71); Canada (60).
- 2007: China (148); USA (147); Canada (43); Taiwan (38).

This distribution reflects the countries hit hardest by SARS (China, Singapore, Taiwan, and Canada) and the role of the USA in leading the public health response. Furthermore, the temporal pattern observed reflects the emergence of a previously unrecognized disease, the discovery of its causative agent, and the control/elimination of disease.



**Fig. 1** SARS SCI articles, 2002–2008

## Publication and citation bibliometrics

Publication metrics are counts of papers published by different entities, and citation metrics are counts of citations to documents published by different entities. The citations were aggregated in all the retrieved SCI papers, the authors, specific documents, years, journals, and countries cited most frequently were identified, and are presented in order of decreasing frequency.

### *Most prolific authors*

The 10 most prolific SARS research authors since the first paper was published (in 2003) and their publication frequency are shown in Table 1. Some of the authors listed collaborated on some of the papers, as shown in Fig. 2. The last column in Table 1 reflects the number of papers on which each author was listed, so there is some double-counting involved.

Five authors (Yuen, Peiris, K. H. Chan, P. K. S. Chan, Sung) account for 65% of the total number of articles ( $N = 497$ ) published by the top 10 most prolific SARS research authors. Nine of the most prolific authors were from China (all from Hong Kong), and the tenth was from Japan. One of the authors was from a research institution, and the other nine were from universities.

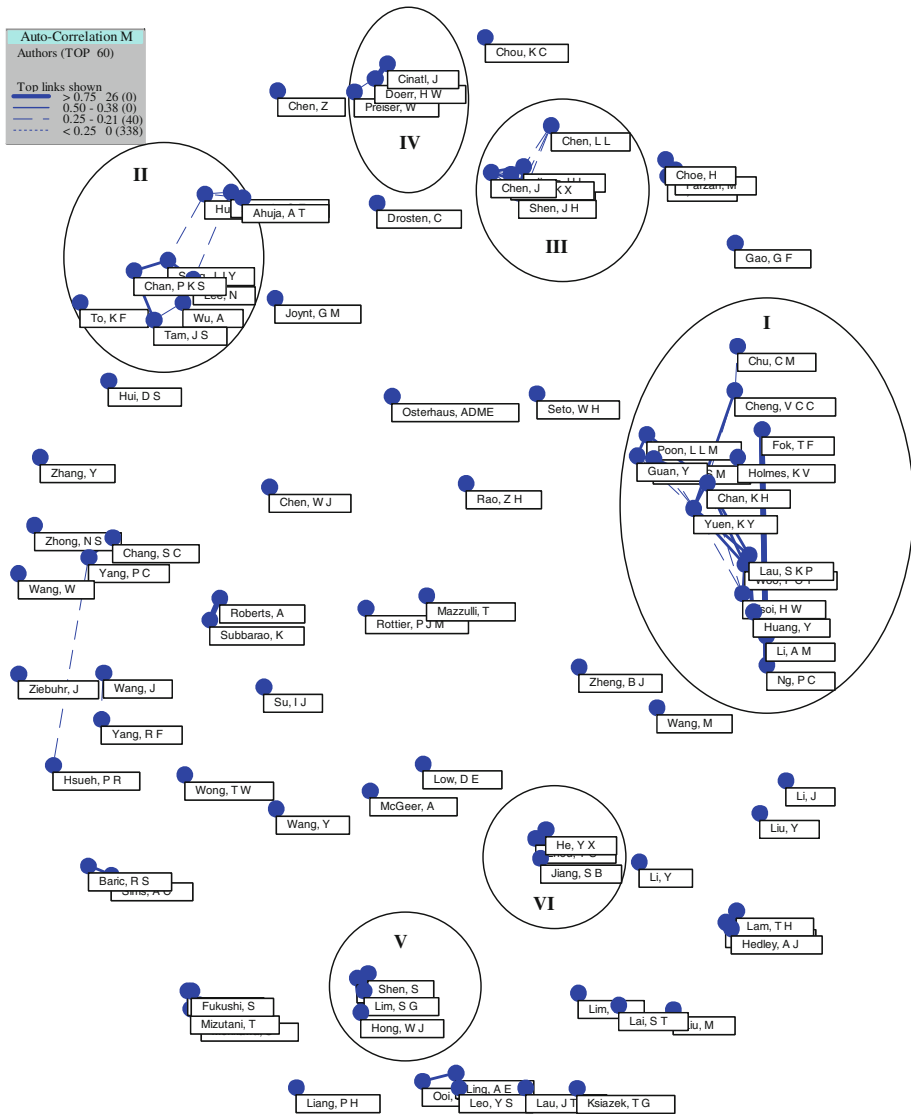
To identify the authors most associated with highly cited SARS papers, the 100 most frequently cited SARS-related documents (as listed in the SCI) were retrieved, and the author frequency extracted. The top 10 central authors are shown in Table 2. Five (50%) of the authors do not appear in the list of most prolific SARS research authors (Table 1). As in Table 1, there is some double-counting due to co-authorship.

The citation data for authors and journals represent citations generated only by the specific records extracted from the SCI database for this study. The data do not represent all the citations received by the papers in those records as they could have received additional citations by papers in other technical disciplines.

Seven of the authors are from China, two are from the Netherlands, and one from the USA. This is the best performance by Chinese authors in absolute terms relative to cited papers in any study we have seen.

**Table 1** Most prolific SARS research authors, 2003–2008

Author	Institution	Country	# REC
Yuen, K. Y.	Univ Hong Kong	China	79
Peiris, J. S. M.	Univ Hong Kong	China	68
Chan, K. H.	Univ Hong Kong	China	60
Chan, P. K. S.	Chinese Univ Hong Kong	China	57
Sung, J. J. Y.	Chinese Univ Hong Kong	China	57
Guan, Y.	Univ Hong Kong	China	40
Li, Y.	Univ Hong Kong	China	36
Poon, L. L. M.	Univ Hong Kong	China	35
Zheng, B. J.	Univ Hong Kong	China	34
Morikawa, S.	Natl Inst Infect Disease	Japan	31



**Fig. 2** Author auto-correlation map

*Author collaborations*

A visual way of displaying collaborating teams as indicated by co-authorship is through an auto-correlation map (an auto-correlation function describes the correlation between a random function and a copy of itself shifted by some ‘lag’ distance. An auto-correlation map of authors shows teams of people who publish together.) The authors are represented as map nodes, and those related are connected by lines of different thicknesses, thicker being stronger.

Figure 2 depicts an author auto-correlation map of the prolific authors, ten of whom are listed in Table 1. Six publishing groups are evident: the largest group (I), (Yuen, Guan,

**Table 2** Authors of most cited SARS papers, 2003–2008

Author	Institution	Country	# REC
Peiris, J. S. M. <sup>a</sup>	Univ Hong Kong	China	11
Yuen, K. Y. <sup>a</sup>	Univ Hong Kong	China	11
Osterhaus, A. D. M. E.	Erasmus Med Col	Netherlands	10
Guan, Y. <sup>a</sup>	Univ Hong Kong	China	9
Chan, K. H. <sup>a</sup>	Univ Hong Kong	China	8
Poon, L. L. M. <sup>a</sup>	Univ Hong Kong	China	8
Lim, W.	Public Health Lab Center	China	7
Subbarao, K.	NIAID/NIH	USA	7
Fouchier, R. A. M	Erasmus Med Col	Netherlands	6
Sung, J. J. Y.	Chinese Univ of Hong Kong	China	6

<sup>a</sup> Author also among top 10 most prolific authors

et al.), is centered around the University of Hong Kong; the second largest group (II), (Chan, Tam, Wu, et al.), is centered around the Chinese University of Hong Kong; the third largest group (III), (Chen, Chen, Shen, et al.), is centered around the Chinese Academy of Science; the fourth largest group (IV), (Cinatl, Doerr, Preiser, et al.), is from the University of Frankfurt Klinikum; the fifth largest group (V), (Shen, Lim, Hong, Tan), is centered around the Institute of Molecular and Cellular Biology, Singapore; and the sixth largest group (VI), (He, Zhang, Zhou, et al.), is from China and the USA. Other than the intra-connection within groups I to VI, there is not a great deal of inter-connection across groups evident from this diagram, based on the threshold values necessary to display linkages.

### *Journals containing most SARS papers*

The nine journals containing the most SARS research papers are listed in Table 3 along with the number of SARS-related papers published between 2003 and 2008. These nine journals contained about  $\frac{1}{4}$  of the total retrieval of 2874 articles.

Two of the journals, *Journal of Virology and Emerging Infectious Diseases*, clearly constitute the first tier of SARS-publishing journals. Many of these journals are highly specialized, and appear quite applied. The biomedical emphases of these journals are medicine (mainly infectious diseases), virology, and microbiology.

**Table 3** Journals publishing most SARS papers, 2003–2008

Journal of Virology	193
Emerging Infectious Diseases	149
Biochemical and Biophysical Research Communications	70
Virology	59
Nidoviruses: Toward Control of SARS and Other Nidovirus Diseases	49
Chinese Medical Journal	44
Proceedings of the National Academy of Sciences of the United States of America	41
Journal of Clinical Microbiology	40
Vaccine	37

**Table 4** Journals publishing most highly cited SARS papers, 2003–2008

Journal	# REC
Lancet	17
Proceedings of the National Academy of Sciences of the United States of America	14
Science	10
New England Journal of Medicine	8
Nature	7
Journal of Virology	5
Nature Medicine	5
Journal of Biological Chemistry	4
Biochemical and Biophysical Research Communications	3
JAMA—Journal of the American Medical Association	3

#### *Journals containing most cited papers*

The top journals in which the 100 most cited SARS papers were published are listed in Table 4. Seventy-six of the 100 most highly cited SARS papers were published in the major medical and multi-disciplinary journals listed in Table 4. No other journal published more than two highly cited papers.

#### *Institutions producing most SARS papers*

The 10 institutions producing the highest numbers of SARS research papers recovered in our search are listed in Table 5.

The 1308 papers in Table 5 represent about 45% of the total 2874 papers. However, the 1308 includes some double-counting because of institutional co-authorship. Six of the top 10 institutions are universities, three are governmental, and one is a hospital. Four are from China, two from the USA, two from Singapore, one from Canada, and one from Taiwan. Five of the institutions stand out in terms of productivity: University of Hong Kong (incl. Queen Mary Hospital); Chinese University of Hong Kong (incl. Prince Wales Hospital);

**Table 5** Institutions producing highest numbers of SARS papers, 2003–2008

Institution	Country	# REC
University of Hong Kong	China	288
Chinese University of Hong Kong	China	251
Chinese Academy of Sciences	China	158
Centers for Disease Control and Prevention	USA	132
National Taiwan University	Taiwan	106
University of Toronto	Canada	92
National University of Singapore	Singapore	83
Peking University	China	73
National Institutes of Health	USA	64
Tan Tock Seng Hospital	Singapore	61



**Table 6** Institutions producing most highly cited SARS papers, 2003–2008

Institution	# HI-CIT	# REC	Ratio
University of Hong Kong	24	288	0.083
Chinese University of Hong Kong	16	251	0.063
National Institutes of Health	9	64	0.140
Centers for Disease Control & Prevention	9	132	0.068
Erasmus Medical Center	8	14	0.571
Harvard University	6	55	0.109
University of Toronto	6	92	0.065
University of Frankfurt	5	23	0.217

Chinese Academy of Science (all institutes); CDC (a public health agency-includes all variants); and National Taiwan University (incl. NTU Hospital).

#### *Institutions producing most highly cited SARS papers*

The institutions producing the most highly cited SARS papers are listed in Table 6.

In Table 6, # HI-CIT is the number of highly cited SARS papers published by the institution, and # REC is the total number of SARS papers published by the institution. The two Hong Kong institutions performed well in terms of absolute numbers of highly cited papers, but when normalized by their total paper production, they are well behind the European institutions of Erasmus Medical Center and University of Frankfurt, and behind the American institutions of Harvard University and NIH as well.

#### *Countries producing the most SARS papers*

The 10 countries producing the most SARS research papers between 2003 and 2008 are listed in Table 7.

These numbers reflect each country's appearance in a paper, and could include some double-counting due to country co-authorship. With the exception of Taiwan and Australia, eight of the top 10 countries participated in the WHO multicenter SARS collaboration and had access to specimens and virus from SARS patients. The first five countries were affected by the SARS outbreak, although to different degrees.

#### *Countries producing most highly cited SARS papers, 2003–2008*

The countries producing the most highly cited SARS papers are listed in Table 8.

The top eight countries were part of the WHO collaborating SARS network. All of these countries had immediate access to specimens and virus from SARS patients.

The ten most highly cited papers are listed in Table 9. Of interest is that all of the top 10 articles were published in 2003. This reflects the fact that SARS was not recognized before October, 2002. The papers can be grouped into several broad categories. The first is the identification of the novel coronavirus associated with SARS and its properties (papers by Ksiazek, Drosten, Rota, Peiris, and Marra). The second group of papers describes the clinical features of SARS (papers by Lee, Peiris, Poutanen, and Tsang). The third group comprises one paper describing viral features that could be exploited for development of

**Table 7** Countries producing most SARS papers, 2003–2008

Country	# REC
Peoples Republic of China	1105
USA	818
Canada	265
Taiwan	256
Singapore	195
UK	164
Germany	162
Japan	108
Netherlands	94
Australia	90

**Table 8** Countries producing most highly cited SARS papers, 2003–2008

Country	# REC
Peoples Republic of China	43
USA	42
Germany	15
Netherlands	14
Canada	10
Singapore	8
England	7
France	6

anti-viral drugs (paper by Li). It is likely that the high number of cites for the top five articles reflects the fact that they are cited in the introduction of most SARS papers.

### Computational linguistics

Thirty-two individual clusters were chosen for the database (2003–2008; Articles and Reviews retrieved from the SCI), and they were agglomerated as shown in the next section.

### *SARS taxonomy*

Table 10 displays the first three levels of the hierarchical taxonomy of the retrieved SARS literature. The first taxonomy level ( $N = 2721$  records with Abstracts) can be sub-divided into two categories: SARS Epidemiology/Clinical Medicine ( $N = 1217$  records) and SARS Biology ( $N = 1504$  records). The SARS Epidemiology/Clinical Medicine category focuses on controlling the spread of infection both in the treatment facility and in the larger local and global population, as well as clinical diagnosis and treatment of SARS. The SARS Biology category focuses on identifying and determining the structure of the SARS coronavirus, and identifying mechanisms and developing drugs for inhibiting the replication of the SARS coronavirus. The boundaries between the two categories are relatively sharp, reflecting the quality of the clustering approach used.

**Table 9** Most cited SARS papers, 2003–2008

Cited author	Time	Journal	Vol	Page	# CIT
Ksiazek, T. G.	2003	New Engl J Med	V348	P1953	903
A novel coronavirus associated with SARS					
Drosten, C.	2003	New Engl J Med	V348	P1967	838
Identification of a novel coronavirus in patients with SARS					
Rota, P. A.	2003	Science	V300	P1394	796
Characterization of a novel coronavirus associated with SARS					
Peiris, J. S. M.	2003	Lancet	V361	P1319	750
Coronavirus as a possible cause of SARS					
Marra, M. A.	2003	Science	V300	P1399	719
The genome sequence of the SARS-associated coronavirus					
Lee, N.	2003	New Engl J Med	V348	P1986	551
Features of a major outbreak of SARS in Hong Kong					
Peiris, J. S. M.	2003	Lancet	V361	P1767	475
Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia					
Li, W. H.	2003	Nature	V426	P450	410
Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus					
Poutanen, S. M.	2003	New Engl J Med	V348	P1995	399
Identification of SARS in Canada					
Tsang, K. W.	2003	New Engl J Med	V348	P1977	354
Clinical features of a cluster of cases of SARS in Hong Kong					

**Table 10** Taxonomy of the SARS literature, 2003–2008

Level 1	Level 2	Level 3
SARS Epidemiology/Clinical Medicine ( <i>N</i> = 1217)	SARS Epidemiology ( <i>N</i> = 820)	Global Epidemic Control ( <i>N</i> = 431) Hospital Epidemic Control ( <i>N</i> = 389)
	SARS Diagnosis and Clinical Treatment ( <i>N</i> = 397)	Clinical Treatment ( <i>N</i> = 294) Clinical Diagnosis ( <i>N</i> = 103)
SARS Biology ( <i>N</i> = 1504)	SARS Coronavirus Structure ( <i>N</i> = 1264)	Coronavirus Genetic Components ( <i>N</i> = 554) Coronavirus Protein Components and Antibodies ( <i>N</i> = 710)
	SARS Drug Development ( <i>N</i> = 240)	Proteolytic Proteinase Inhibitors ( <i>N</i> = 118) Viral Inhibiting Drugs ( <i>N</i> = 122)

For the second level taxonomy, each first level category is divided into two sub-categories. SARS Epidemiology/Clinical Medicine is divided into SARS Epidemiology (*N* = 820 records) and SARS Diagnosis and Clinical Treatment (*N* = 397 records), while SARS Biology is divided into SARS Coronavirus Structure (*N* = 1264 records) and SARS Drug Development (*N* = 240 records). The SARS Epidemiology category focuses on controlling the spread of infection both in health care facilities as well as in the larger local and global settings, while the SARS Diagnosis and Clinical Treatment category focuses on

clinical diagnosis and treatment for SARS. The SARS Coronavirus Structure category focuses on determining the structure of the SARS coronavirus, while the SARS Drug Development category focuses on developing drugs for inhibiting the replication of the SARS coronavirus.

The second level categories are further sub-divided to form eight-third level categories. This will be the final level discussed. The third level is treated as a flat taxonomy. The category heading (in bolded italics) is followed by the category summary metrics (prolific Authors, Countries, Institutions), which are followed by the component cluster themes, bulletized.

The first four Level 3 categories are under the Level 1 SARS Epidemiology/Clinical Medicine category.

Category 1, *Global Epidemic Control* ( $N = 431$  records):

(Country: USA 157; Peoples R China 97; Canada 50; England 41; Institution: Univ Hong Kong 25; Ctr Dis Control & Prevent 14; WHO 13; Chinese Univ Hong Kong 12; Univ London Imperial Coll Sci Technol & Med 11; Univ British Columbia 11; Chinese Acad Sci 11)

- Emerging infectious disease ( $N = 70$  records)
- Pandemic influenza ( $N = 67$  records)
- Public health monitoring ( $N = 140$  records)
- SARS quarantine ( $N = 84$  records)
- Modeling SARS epidemiology ( $N = 70$  records)

Category 2, *Hospital Epidemic Control* ( $N = 389$  records):

(Country: Peoples R China 160; USA 73; Canada 63; Taiwan 58; Singapore 48; Institution: Chinese Univ Hong Kong 71; Univ Hong Kong 54; Univ Toronto 37)

- SARS hospital airborne transmission ( $N = 65$  records)
- SARS hospital infection control ( $N = 149$  records)
- SARS health care worker protection ( $N = 78$  records)
- SARS nursing impacts ( $N = 37$  records)
- Psychological responses to SARS ( $N = 60$  records)

Category 3, *Clinical Treatment* ( $N = 294$  records):

(Country: Peoples R China 151; Taiwan 51; USA 39; Singapore 26; Canada 25; Institution: Chinese Univ Hong Kong 50; Univ Hong Kong 41; Princess Margaret Hosp 28)

- SARS clinical outcomes ( $N = 135$  records)
- SARS clinical features and diagnosis ( $N = 159$  records)

Category 4, *Clinical Diagnosis* ( $N = 103$  records):

(Country: Peoples R China 60; Singapore 14; Taiwan 13; Canada 12; USA 10; Institution: Chinese Univ Hong Kong 23; Univ Hong Kong 13; Tan Tock Seng Hosp 10; Princess Margaret Hosp 8)

- Chest radiographs ( $N = 60$  records)
- SARS pulmonary pathology ( $N = 43$  records)

The last four Level 3 categories are under the Level 1 SARS Biology category.

Category 5, *Coronavirus Genetic Components* ( $N = 554$  records):

(Country: Peoples R China 216; USA 182; Germany 51; Netherlands 42; Canada 40; Institution: Univ Hong Kong 56; Chinese Acad Sci 38; Chinese Univ Hong Kong 27; Ctr Dis Control & Prevent 17)

- SARS interferon treatments ( $N = 54$  records)
- SARS lung pathology ( $N = 126$  records)
- SARS coronavirus detection using PCR ( $N = 105$  records)
- Human coronavirus NL63 ( $N = 25$  records)
- SARS coronavirus natural reservoirs ( $N = 47$  records)
- Genome sequencing of SARS coronavirus ( $N = 97$  records)
- RNA virus ( $N = 100$  records)

Category 6, *Coronavirus Protein Components and Antibodies* ( $N = 710$  records):

(Country: Peoples R China 299; USA 226; Singapore 58; Taiwan 49; Japan 49; Institution: Chinese Acad Sci 75; Univ Hong Kong 50; NIAID 29; Inst Mol & Cell Biol 27; Harvard Univ 26)

- SARS coronavirus vaccine ( $N = 143$  records)
- SARS coronavirus nucleocapsid protein epitopes ( $N = 86$  records)
- SARS coronavirus antibody detection ( $N = 83$  records)
- SARS coronavirus protein structure ( $N = 167$  records)
- SARS coronavirus-induced Vero E6 cell apoptosis ( $N = 77$  records)
- Angiotensin-converting enzyme as SARS-CoV receptor ( $N = 76$  records)
- SARS-CoV spike protein fusion peptide ( $N = 78$  records)

Category 7, *Protease/Proteinase Inhibitors* ( $N = 118$  records):

(Country: Peoples R China 50; USA 27; Taiwan 24; Canada 17; Germany 11; Institution: Chinese Acad Sci 21; Acad Sinica 12; Univ British Columbia 8; Peking Univ 8; Natl Taiwan Univ 8)

- SARS coronavirus proteinase inhibitors ( $N = 57$  records)
- SARS coronavirus main proteinase 3CL(pro) ( $N = 61$  records)

Category 8, *Viral Inhibiting Drugs* ( $N = 122$  records):

(Country: USA 40; Peoples R China 34; Taiwan 11; Belgium 11; Japan 10; Italy 10; Germany 9; Institution: Univ Hong Kong 8; Katholieke Univ Leuven 8; Acad Sinica 8; Tianjin Normal Univ 6; Gordon Life Sci Inst 6)

Categories 1 and 2 are epidemiology-related. Category 1 focuses on the broader epidemiological issues, such as classes of emerging infectious diseases, surveillance, monitoring, and quarantine during outbreaks, and modeling epidemiologic dynamics. Overall, China produced about 35% more papers than the USA; however, in this category the USA outproduced China by about 60%. Canada, which overall had 25% of the papers of China, had about half of the Chinese papers in this category. Also overall, the Chinese University of Hong Kong had almost as many papers as the University of Hong Kong; but, in this category it had about half as many.

Category 2 focused on attempts to control the spread of infection within the health care setting, and to identify the effects of SARS on the physical and mental health of the care givers. In this category, unlike Category 1, China had more than twice as many papers as the USA, perhaps reflecting the urgency and much more operational experience with SARS relative to the USA. Also, in this case, the Chinese University of Hong Kong had almost 50% more papers than the University of Hong Kong.

Categories 3 and 4 reflect the clinical aspects of SARS. Category 3 addresses clinical treatments, emphasizing outcomes and clinical features. In this category, China outproduced the USA by a factor of four, again reflecting the much more extensive experience in treating patients with SARS. Category 4 addresses clinical diagnosis, emphasizing chest

radiographs and pulmonary pathology. China outproduced the USA by a factor of six, reflecting again its greater experience. While the numbers of records in Category 4 are relatively small, the Chinese University of Hong Kong outproduced the University of Hong Kong by almost a factor of two.

Categories 5 and 6 reflect various aspects of the SARS coronavirus and its pathogenesis. Category 5 is relatively broad, addressing the genetic components of the coronavirus as well as the structural components. The striking feature of this category is that University of Hong Kong now has double the paper output of Chinese University of Hong Kong.

Category 6 reflects the protein components of SARS coronavirus and their role in pathogenesis, host immune response, and as potential vaccine candidates. Here, the Chinese Academy of Science has 50% more papers than the University of Hong Kong, and the Chinese University of Hong Kong is not even among the top five producers.

Categories 7 and 8 reflect issues related to drug development. Category 7 addresses protease/proteinase inhibitors. The Hong Kong universities don't even show up among the first five, the leaders being Chinese Academy of Science, followed by *Academica Sinica*. Category 8 addresses virus inhibiting drugs. In this category, Belgium, which had slightly over 2% of China's papers overall, has one-third of China's paper production, with the main Belgian institution being Catholic University of Louvaine.

To summarize, the Chinese University of Hong Kong played the strongest role in the patient care and epidemiology issues, while the University of Hong Kong and Chinese Academy of Sciences played a leading role in the biological research output. Among USA institutions, CDC played a strong role in the epidemiology of SARS and in the characterization of the SARS coronavirus. NIAID and Harvard played an important role in elucidating the role of viral proteins in the immune response and as potential vaccine candidates.

## Summary and conclusions

SARS was the first major pandemic of the twenty first century. The nations involved were essentially caught unprepared, since there are few efforts devoted to predicting new pathogenic forms such as biowarfare agents or emerging infectious diseases (Swanson et al. 2001). Nevertheless, multilateral collaboration was rapidly established, as exemplified by the WHO-organized multicenter collaborative network. As the literature bibliometrics results show, PRC played a major role in all aspects of diagnosing, treating, and researching the disease. In fact, PRC had a higher fraction of most highly cited papers in this study than in any we have seen before. Within PRC, the researchers at the Hong Kong institutions played a major role in significant advances across the SARS spectrum.

According to recent surveys, none of the treatments used during the pandemic were effective. 8000 people were reported to have contracted SARS, and about 10% of those died. Given the large numbers of contacts, why didn't far more people contract SARS? And, even though ten percent died, how did the other 90% survive, given the lack of effective treatment?

We believe there is a message in the fact that these numbers were not substantially larger than they were. We will be using these results in the third part of the study, identifying potential discovery (treatments and preventatives) related to SARS.

## References

- Davidse, R. J., & Van Raan, A. F. J. (1997). Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics. *Scientometrics*, *40*(2), 171–193.
- Feng, Y. J., & Gao, G. F. (2007). Towards our understanding of SARS-CoV, an emerging and devastating but quickly conquered virus. *Comparative Immunology, Microbiology and Infectious Diseases*, *30*(5–6), 309–327.
- Garfield, E. (1985). History of citation indexes for chemistry—a brief review. *JCICS*, *25*(3), 170–174.
- Goldman, J. A., Chu, W. W., Parker, D. S., & Goldman, R. M. (1999). Term domain distribution analysis: A data mining tool for text databases. *Methods of Information in Medicine*, *38*, 96–101.
- Greengrass, E. (1997). Information retrieval: An overview. National Security Agency. TR-R52-02-96.
- Hao, P., Chen, M., Zhang, G. Q., He, W. Z., & Li, Y. X. (2006). Bioinformatics research on the SARS coronavirus (SARS\_CoV) in China. *Current Pharmaceutical Design*, *12*(35), 4565–4572.
- Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of ACL 99, the 37th annual meeting of the association for computational linguistics*, June 20–26, 1999. University of Maryland.
- Janies, D., Habib, F., Alexandrov, B., Hill, A., & Pol, D. (2008). Evolution of genomes, host shifts and the geographic spread of SARS-CoV and related coronaviruses. *Cladistics*, *24*(2), 111–130.
- Karypis, G. (2004). CLUTO—A clustering toolkit. <http://www.cs.umn.edu/cluto>.
- Kostoff, R. N. (2003). Bilateral asymmetry prediction. *Medical Hypotheses*, *61*(2), 265–266.
- Kostoff, R. N. (2008). Literature-related discovery: Introduction and background. In R. N. Kostoff (Ed.), Special issue on literature-related discovery. *Technological forecasting and social change*, *75*(2), 165–185.
- Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., & Humenik, J. (2000). Fullerene roadmaps using bibliometrics and Database Tomography. *Journal of Chemical Information and Computer Science*, *40*(1), 19–39.
- Kostoff, R. N., Briggs, M., Rushenberg, R., Bowles, C. A., & Pecht, M. (2006). *The structure and infrastructure of Chinese science and technology*. DTIC Technical report number ADA443315. Fort Belvoir, VA: Defense Technical Information Center. <http://www.dtic.mil/>.
- Kostoff, R. N., Briggs, M. B., Solka, J. A., Rushenberg, R. L. (2008). Literature-related discovery: Methodology. In R. N. Kostoff (Ed.), *Special issue on literature-related discovery. Technological Forecasting and Social Change*, *75*(2), 186–202.
- Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., & Humenik, J. A. (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, *52*(13), 1148–1156.
- Kostoff, R. N., Eberhart, H. J., & Toothman, D. R. (1997). Database Tomography for information retrieval. *Journal of Information Science*, *23*(4), 301–311.
- Kostoff, R. N., Morse, S. A., & Oncu, S. (2007). The seminal literature of anthrax research. *Critical Reviews in Microbiology*, *33*(3), 171–181.
- Kostoff, R. N., Shlesinger, M. F., & Malpohl, G. (2004a). Fractals roadmaps using bibliometrics and Database Tomography. *Fractals*, *12*(1), 1–16.
- Kostoff, R. N., Shlesinger, M. F., & Tshitaya, R. (2004b). Nonlinear dynamics roadmaps using bibliometrics and Database Tomography. *International Journal of Bifurcation and Chaos*, *14*(1), 61–92.
- Losiewicz, P., Oard, D., & Kostoff, R. N. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, *15*, 99–119.
- Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C-637. National Science Foundation. 1976. Contract NSF C-627. NTIS accession no. PB252339/AS.
- Narin, F., Olivastro, D., & Stevens, K. A. (1994). Bibliometrics theory, practice and problems. *Evaluation Review*, *18*(1), 65–76.
- Schubert, A., Glanzel, W., & Braun, T. (1987). Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics*, *12*(5–6), 267–291.
- Swanson, D. R. (1986). Fish Oil, Raynauds Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18.
- Swanson, D. R., Smalheiser, N. R., & Bookstein, A. (2001). Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology*, *52*(10), 797–812.
- Zhang, Z. B. (2007). The outbreak pattern of SARS cases in China as revealed by a mathematical model. *Ecological Modelling*, *204*(3–4), 420–426.
- Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, *55*(3), 311–331.