

Mining Candidate Viruses as Potential Bio-terrorism Weapons from Biomedical Literature

Xiaohua Hu¹, Illhoi Yoo¹, Peter Rumm², and Michael Atwood¹

¹ College of Information Science and Technology, Drexel University,
Philadelphia, PA 19104

{thu, michael.atwood}@cis.drexel.edu
{iy28, pdr26}@cis.drexel.edu

² School of Public Health, Drexel University, Philadelphia, PA 19104

Abstract. In this paper we present a semantic-based data mining approach to identify candidate viruses as potential bio-terrorism weapons from biomedical literature. We first identify all the possible properties of viruses as search key words based on Geissler's 13 criteria; the identified properties are then defined using MeSH terms. Then, we assign each property an importance weight based on domain experts' judgment. After generating all the possible valid combinations of the properties, we search the biomedical literature, retrieving all the relevant documents. Next our method extracts virus names from the downloaded documents for each search keyword and identifies the novel connection of the virus according to these 4 properties. If a virus is found in the different document sets obtained by several search keywords, the virus should be considered as suspicious and treated as candidate viruses for bio-terrorism. Our findings are intended as a guide to the virus literature to support further studies that might then lead to appropriate defense and public health measures.

1 Introduction

The threat of bio-terrorism is real. The anthrax mail attack in October, 2001 terrorism caused 23 cases of anthrax-related illness and 5 deaths. The threat of the use of biological weapons against public is more acute than any time in U.S. history due to the widespread availability of biological/chemical agents, widespread knowledge of production methodologies, and potential dissemination devices. Therefore, the discovery of additional viruses as bio-terrorism weapon and preparedness for this threat is seemingly vital to the public health and home land security.

Because it is very difficult for laypeople to diagnose and recognize most of the diseases caused by biological weapons, we need surveillance systems to keep an eye on potential uses of such biological weapons [2]. Before initiating such systems, we should identify what biological agents could be used as biological weapons. Geissler identified and summarized 13 criteria (shown in Table 1) to identify biological warfare agents as viruses [6]. Based on the criteria, he compiled 21 viruses. Figure 1 lists the 21 virus names in MeSH terms. The viruses in Figure 1 meet some of the criteria described in Table 1.

Table 1. Geissler's 13 Criteria for Viruses

1	The agent should consistently produce a given effect: death or disease.
2	The concentration of the agent needed to cause death or disease the infective dose should be low.
3	The agent should be highly contagious.
4	The agent should have a short and predictable incubation time from exposure to onset of the disease symptoms.
5	The target population should have little or no natural or acquired immunity or resistance to the agent.
6	Prophylaxis against the agent should not be available to the target population.
7	The agent should be difficult to identify in the target population, and little or no treatment for the disease caused by the agent should be available.
8	The aggressor should have means to protect his own forces and population against the agent clandestinely.
9	The agent should be amenable to economical mass production.
10	The agent should be reasonably robust and stable under production and storage conditions, in munitions and during transportation. Storage methods should be available that prevent gross decline of the agent's activity.
11	The agent should be capable of efficient dissemination. If it cannot be delivered via an aerosol, living vectors (e.g. fleas, mosquitoes or ticks) should be available for dispersal in some form of infected substrate.
12	The agent should be stable during dissemination. If it is to be delivered via an aerosol, it must survive and remain stable in air until it reaches the target population.
13	After delivery, the agent should have low persistence, surviving only for a short time, thereby allowing a prompt occupation of the attacked area by the aggressor's troops

▪ Hemorrhagic Fever Virus, Crimean-Congo	▪ Encephalitis Virus, Eastern Equine	
▪ Lymphocytic choriomeningitis virus	▪ Encephalitis Virus, Japanese	
▪ Encephalitis Virus, Venezuelan Equine	▪ Encephalitis Viruses, Tick-Borne	
▪ Encephalitis Virus, Western Equine	▪ Encephalitis Virus, St. Louis	
▪ Arenaviruses, New World	▪ Chikungunya virus	▪ Hepatitis A virus
▪ Marburg-like Viruses	▪ Dengue Virus	▪ Orthomyxoviridae
▪ Rift Valley fever virus	▪ Ebola-like Viruses	▪ Junin virus
▪ Yellow fever virus	▪ Hantaan virus	▪ Lassa virus
		▪ Variola virus

Fig. 1. Geissler's 21 Viruses

Based on the criteria, government agencies such as CDC and the Department of Homeland Security compile and monitor viruses which are known to be dangerous in

bio-terrorism. One problem of this approach is that the list is compiled manually, requiring extensive specialized human resources and time. Because the viruses are evolving through mutations, biological or chemical change, some biological substances have the potential to turn into deadly virus through chemical/genetic/biological reaction, there should be an automatic approach to keep track of existing suspicious viruses and to discover new viruses as potential weapons. We expect that it would be very useful to identify those biological substances and take precaution actions or measurements.

2 Related Works

The problem of mining implicit knowledge/information from biomedical literature was exemplified by Dr. Swanson's pioneering work on Raynaud disease/fish-oil discovery in 1986 [11]. Back then, the Raynaud disease had no known cause or cure, and the goal of his literature-based discovery was to uncover novel suggestions for how Raynaud disease might be caused, and how it might be treated. He found from biomedical literature that Raynaud disease is a peripheral circulatory disorder aggravated by high platelet aggregation, high blood viscosity and vasoconstriction. In another separate set of literature on fish oils, he found out the ingestion of fish oil can reduce these phenomena. But no single article from both sets in the biomedical literature mentions Raynaud and fish oil together in 1986. Putting these two separate literatures together, Swanson hypothesized that fish oil may be beneficial to people suffering from Raynaud disease [11][12]. This novel hypothesis was later clinically confirmed by DiGiacomo in 1989 [4]. Later on [10] Dr. Swanson extended his methods to search literature for potential virus. But the biggest limitation of his methods is that, only 3 properties/criteria of a virus are used as search key word and the semantic information is ignored in the search procedure. In this paper, we present a novel biomedical literature mining algorithms based on this philosophy with significant extensions. Our objective is to extend the existing known virus list compiled by CDC to other viruses that might have similar characteristics. We hypothesize, therefore, that viruses that have been researched with respect to the characteristics possessed by existing viruses are leading candidates for extending the virus lists. Our findings are intended as a guide to the virus literature to support further studies that might then lead to appropriate defense and public health measures.

3 Method

We propose an automated, semantic-based data mining system to identify viruses that can be used as potential weapons in bio-terrorism. Following the criteria established by Geissler and the similar ideas used by Swanson [10], in the mining procedure, we consider many important properties of the virus such as *the genetic aspects of virulence; airborne transmission of viral disease; and stability of viruses in air or aerosol mixtures* etc.. Our objective is to identify which viruses have been investigated with respect to these properties. The main assumption of the proposed approach is that the more criteria are met by a virus, the more suspicious the virus is a potential candidate

for bio-terrorism. In other words, if a virus is commonly found in the different document sets searched by several search keywords, the virus should be considered as suspicious.

We introduce an automated semantic-based search system, called Combinational Search based Virus Seeker (CSbVS), to identify viruses that can be used as potential weapons in bio-terrorism. The method is based on Dr. Swanson's method with the following enhancements:

- (1) Search keywords (SK) are more complete based on Dr. Geissler's 13 criteria.
- (2) The importance of search key words are reflected by different weights based on the properties of the virus.
- (3) In [10], only 3 properties/criteria of a virus are used as search key word, we consider all the meaningful combinations of the properties/criteria of the virus. And different search keywords have different weight; if a virus is found to meet the criteria in many search keywords, the virus is more suspicious. Therefore, the result is more reliable. Each virus has its own score so that the viruses can be ranked while Swanson just listed the viruses without any ranking.

In order to find all the suspicious viruses in the biomedical literature, we first identify all the possible properties of viruses as search key words based on Geissler's 13 criteria; the identified terms are then defined in MeSH terms (a biomedical ontology developed by the National Library of Medicine, <http://www.nlm.nih.gov/mesh/meshhome.html>). These properties are shown in Figure 2. Then, we assign each

▪ "Virulence"[MeSH]	
▪ "Disease Outbreaks"[MeSH]	
▪ "Viral Nonstructural Proteins"[MeSH]	
▪ "Cross Reactions"[MeSH]	
▪ "Mutation"[MeSH] AND "Virus Replication"[MeSH]	
▪ "Insect Vectors"[MeSH]	
▪ "severe acute"	▪ fever
▪ cause OR causing	▪ hemorrhagic
▪ mortality	▪ infect OR infecting
▪ death AND disease	▪ mosquito-borne
▪ encephalitis OR encephalomyelitis	▪ transmission OR transmit
▪ epidemics OR epidemiologically	▪ survive
▪ etiologic	▪ viability OR viable
▪ fatal	▪ airborne
▪ febrile	

Fig. 2. The Properties/Criteria of Suspicious Viruses

property an importance weight based on domain experts' judgment. For example, it is believed that "virulence" as MeSH term is much more important than "cause" in searching the potential virus. Therefore, for each search keyword, a weight is given based on the importance; this is the domain knowledge, which may lead to better results to identify suspicious virus. After generating all the possible valid combinations

of the properties, CSbVS performs the searches for each combination, retrieving all the relevant documents. Each combination has its importance, which is the sum of the weights of the key words used in the combination. Next CSbVS extracts virus names from the downloaded documents for each search keyword and identifies the novel connection of the virus with these properties. The viruses, as a result of each search combination, have the same importance as the search combination. Based on the importance, all the viruses are ranked. Figure 3 shows the data flow of CSbVS.

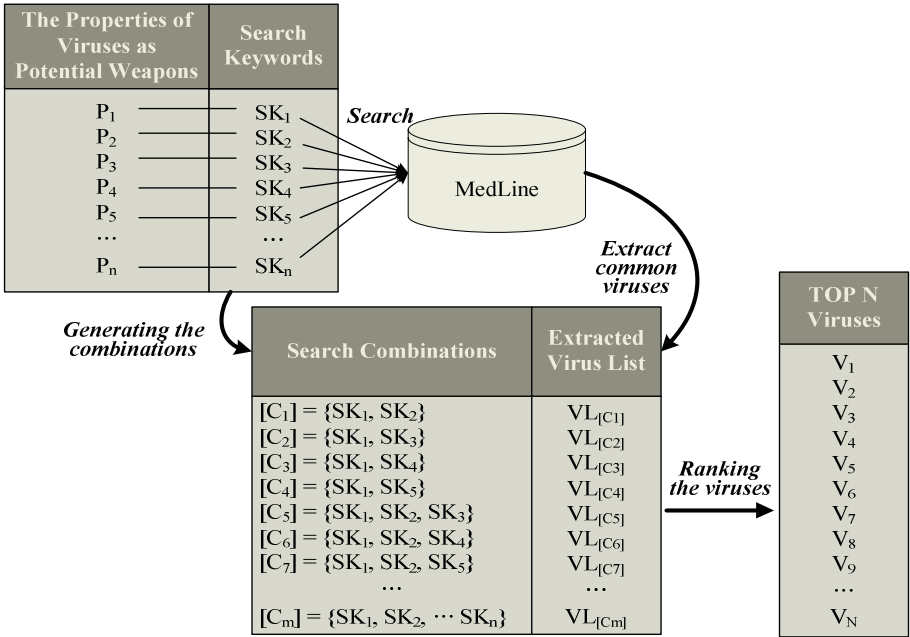


Fig. 3. The Data Flow of CSbVS

Figure 4 shows an example of the combinational search. Each circle (e.g., Virus_{sk1}) indicates a virus list from the documents by a search keyword (e.g., SK1). Each

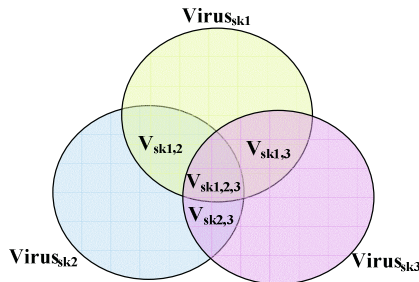


Fig. 4. An Example of Combinational Search

intersection (e.g., $V_{sk1,2}$) contains the viruses that are commonly found in the original document sets. Therefore, we can guess the viruses in $V_{sk1,2,3}$ are more suspicious than the viruses in $V_{sk1,2}$.

Input: Search keywords with their weights

All virus names in MeSH terms

Output: the top N viruses

Procedural

STEP 1: Generating all the possible valid combinations of the search keywords.

STEP 2: Searching every keyword and extracting virus names based on the virus category in the MeSH hierarchy in the downloaded documents for each search keyword

STEP 3: Finding common viruses in various search combination

STEP 4: Accumulate the scores of the common viruses with the sum of the weights of the search keywords involved

STEP 5: Sorting all viruses based on their accumulated scores in descending order.

Fig. 5. The Algorithm of CSbVS

STEP 1: Generating all the possible valid combinations of search keywords. For example, if there are 4 search keywords (e.g., A, B, C, D), all the possible valid combinations used in the approach are the following. {AB, ABC, ABCD, AC, AD, BCD, BD, CD, ABD}

Here we assume that a virus has to meet at least two criteria to be considered as a potential virus for bio-terrorism. The combinations that consist of only a single criterion are not considered.

STEP 2: Searching every keyword against Medline (PubMed) and download the documents relevant to each search keyword. For better recall and precision, we included “Viruses” and “Human” as MeSH terms into the combinational search. For example, for the “Virulence” search keyword, the complete search keyword against PubMed is the following

"Virulence"[MeSH] AND "Viruses"[MAJOR] AND "Human"[MeSH]

After downloading the relevant documents, the system extracts virus names from these documents for each search keyword; the targets of the extraction are *Major-Topic* virus names assigned to Medline articles. In order to identify virus names, we collected all the MeSHs in B04 category (“Viruses”) of MeSH Categories that are shown Appendix 1; the total number of MeSH terms in the category is 487.

STEP 3: Finding common viruses in each combination. If a combination consists of A, B & C as search key, the virus list contains the viruses that are “commonly” found in every document set by the search key as shown in Figure 6.

STEP 4: Accumulate the scores of the common viruses based on the weights of the search keywords involved.

STEP 5: Sorting all viruses based on their accumulated scores in descending order. Finally the top N viruses are the output.

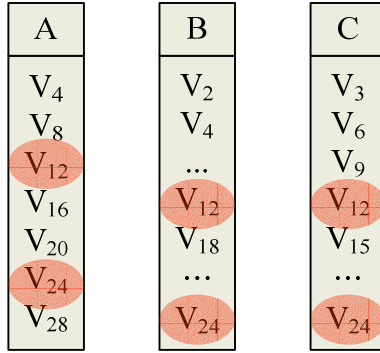


Fig. 6. Viruses commonly found in A, B, and C document sets

4 Experimental Results

In our experiment, first of all, a weight (1 to 5) is carefully assigned to each property based on domain expert’s opinion. Then, CSbVS download the documents relevant to each keyword from Medline and extracted virus names for each combination of search keywords. Table 2 shows the number of documents for each search keyword. After searching every search keyword against Medline and generating all the possible valid combination of search keywords, common virus names are extracted for each combination search. Finally, based on the importance of the combinational search, all the viruses are ranked. Table 3 shows the top 143 suspicious viruses; which included all the 21 viruses identified by Geissler (marked in bold.)

Table 2. The Search Keywords and the number of Documents by them

Search Keywords	# of Doc.
Virulence[MeSH]	1455
Disease Outbreaks[MeSH]	3141
Viral Nonstructural Proteins[MeSH]	1262
Cross Reactions[MeSH]	1559
("Mutation"[MeSH] AND "Virus Replication"[MeSH])	1742
Insect Vectors[MeSH]	413
severe acute	487
(cause OR causing)	5907
mortality	3279
(death AND disease)	1052
(encephalitis OR encephalomyelitis)	4398
epidemics OR epidemiologically)	17825
etiologic	988
fatal	1372
febrile	746

Table 2. (Continued...)

Search Keywords	# of Doc.
fever	3629
hemorrhagic	1412
(infect OR infecting)	2883
mosquito-borne	98
(transmission OR transmit)	11166
survive	196
(viability OR viable)	1081
airborne	61
total	66152

Table 3. The top 143 suspicious viruses

Ranking	Virus Name in MeSH	Ranking	Virus Name in MeSH
1	Hepacivirus	26	Herpesvirus 6, Human
2	West Nile virus	27	Rotavirus
3	<i>Dengue Virus</i>	28	Sindbis Virus
4	<i>Encephalitis Viruses, Tick-Borne</i>	29	Respirovirus
5	Hantavirus	30	Flavivirus
6	Bunyaviridae	31	<i>Yellow fever virus</i>
7	Vaccinia virus	32	Arboviruses
8	Herpesvirus 3, Human	33	Encephalitis Viruses
9	Enterovirus	34	Herpesvirus 8, Human
10	Respiratory Syncytial Viruses	35	<i>Orthomyxoviridae</i>
11	Adenoviruses, Human	36	Polioviruses
12	Cytomegalovirus	37	Bunyamwera virus
13	Adenoviridae	38	Rabies virus
14	Influenza A Virus, Human	39	Influenza B virus
15	Herpesvirus 4, Human	40	HIV
16	Enterovirus B, Human	41	Respiratory Syncytial Virus, Human
17	Influenza A virus	42	Measles virus
18	Herpesvirus 2, Human	43	Alphavirus
19	Herpesviridae	44	<i>Encephalitis Virus, Japanese</i>
20	Herpesvirus 1, Human	45	Deltaretrovirus
21	Simplexvirus	46	Parvovirus B19, Human
22	HIV-1	47	Picornaviridae
23	<i>Ebola-like Viruses</i>	48	RNA Viruses
24	Human T-lymphotropic virus 1	49	Reassortant Viruses
25	<i>Encephalitis Virus, Venezuelan Equine</i>	50	SARS Virus

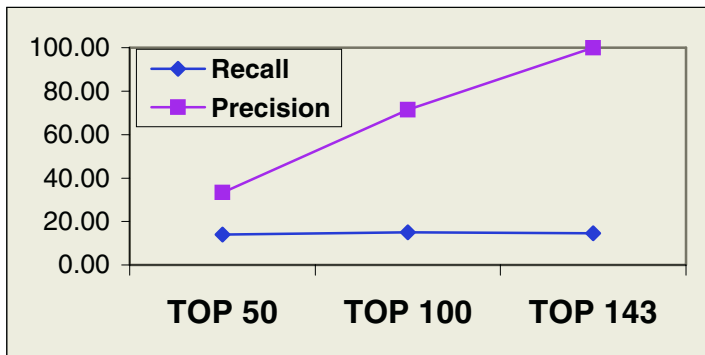
Table 3. (Continued...)

Ranking	Virus Name in MeSH	Ranking	Virus Name in MeSH
51	<i>Hantaan virus</i>	88	Bacteriophages
52	Influenza A Virus, Avian	89	Parvoviridae
53	Flaviviridae	90	Newcastle disease virus
54	Retroviridae	91	Endogenous Retroviruses
55	<i>Rift Valley fever virus</i>	92	Rubella virus
56	Norovirus	93	Papillomavirus
57	Vesicular stomatitis-Indiana virus	94	Coliphages
58	Hepatitis A Virus, Human	95	Semliki forest virus
59	Hepatovirus	96	HIV-2
60	Virion	97	Hepatitis Viruses
61	Mumps virus	98	<i>Lymphocytic choriomeningitis virus</i>
62	Influenza A Virus, Porcine	99	Proviruses
63	Poliovirus	100	Coronaviridae
64	Rhinovirus	101	Caliciviridae
65	Morbillivirus	102	<i>Encephalitis Virus, Eastern Equine</i>
66	Papillomavirus, Human	103	Herpesvirus 7, Human
67	SIV	104	Salmonella Phages
68	Paramyxovirinae	105	Lentivirus
69	<i>Hemorrhagic Fever Virus, Crimean-Congo</i>	106	Lyssavirus
70	Parainfluenza Virus 3, Human	107	Echovirus 9
71	Poxviridae	108	Parainfluenza Virus 1, Human
72	Hepatitis Delta Virus	109	Distemper Virus, Canine
73	<i>Arenaviruses, New World</i>	110	Encephalomyocarditis virus
74	<i>Encephalitis Virus, St. Louis</i>	111	Simian virus 40
75	Astrovirus	112	Metapneumovirus
76	Arenaviridae	113	Norwalk virus
77	Hepatitis E virus	114	<i>Chikungunya virus</i>
78	Oncogenic Viruses	115	Aphthovirus
79	JC Virus	116	Ross river virus
80	DNA Viruses	117	Viruses, Unclassified
81	<i>Lassa virus</i>	118	SSPE Virus
82	<i>Marburg-like Viruses</i>	119	Filoviridae
83	Rhabdoviridae	120	Monkeypox virus
84	Reoviridae	121	Herpesvirus 1, Cercopithecine
85	Coronavirus	122	Encephalitis Virus, Murray Valley
86	Defective Viruses	123	Theilovirus
87	Polyomavirus		

Table 3. (Continued...)

Ranking	Virus Name in MeSH	Ranking	Virus Name in MeSH
124	BK Virus	133	<i>Hepatitis A virus</i>
125	Orthopoxvirus	134	Papillomaviridae
126	<i>Variola virus</i>	135	Echovirus 6, Human
127	Paramyxoviridae	136	Leukemia Virus, Murine
128	Murine hepatitis virus	137	Phlebovirus
129	Borna disease virus	138	Muromegalovirus
130	Transfusion-Transmitted Virus	139	Baculoviridae
131	<i>Encephalitis Virus, Western Equine</i>	140	Parvovirus
132	Dependovirus	141	Coronavirus OC43, Human
		142	Herpesvirus 1, Suid

Although Geissler's 21 viruses, compiled in 1986, would not be the full list of the viruses used as potential weapons at present, we compare our Top 50, 100 & 143 Viruses with Geissler's 21 viruses as a golden standard in terms of recall and precision; all of the Geissler's 21 viruses are found within our top 143 viruses. As Figure 7 shows, the recalls are consistent for the three groups. In other words, Geissler's 21 viruses are equally distributed in our Top 143 virus list. It is very important to note that our system is able to find "West Nile virus" and "SARS Virus", and ranks them in 2nd and 50th respectively.

**Fig. 7.** Recalls and Precisions for Top 50, 100 and 143

5 Potential Significance for Public Health and Homeland Security

As such a list shows there are many potential viral threats that could affect the health of the public on a wide scale if disseminated effectively. This situation is worrisome to public health officials who are concerned that the public health system might not yet be prepared fully for such a crisis as a release of a viral agent in the U.S. population. Certainly, there have been steps made in laboratory preparedness and public

health preparedness to identify such threats but potential gaps remain [1][2][6][7][8]. These viruses vary in their biological capability to survive, replicate, and be infective. The U.S. Department of Health and Human Services Agency for Health Care Research and Quality working with University of Alabama has recently put out a list of what they think are the most probable public health biological threats with following caveats [13]:

The U.S. public health system and primary health-care providers must be prepared to address varied biological agents, including pathogens that are rarely seen in the United States. High-priority agents include organisms that pose a risk to national security because they

- can be easily disseminated or transmitted person-to-person
- cause high mortality, with potential for major public health impact
- might cause public panic and social disruption; and require special action for public health preparedness

The Category A viral agents that most fit this bill currently according to AHRQ, the CDC, and the University of Alabama (and other experts) are: Smallpox, viral hemorrhagic agents (there are many of these – see below), SARS and Monkeypox.

Besides Smallpox, which is known to have stores in secured locations in the U.S. and Russia but may be in other sites [3], most of the literature currently focuses on the usage of hemorrhagic viruses as the most probably viral bio-terrorism threats.

Other viruses which our federal government conceives be used as terrorist weapons on a population scale, but not considered as great a threat as the filoviruses include:

Arenaviruses: Lassa Fever (Africa) and the New World Hemorrhagic Fevers - Bolivian Hemorrhagic Fever (BHF, Machupo virus), Argentine Hemorrhagic Fever (AHF, Junin virus), Venezuelan Hemorrhagic Fever (Guanarito virus), and Brazilian Hemorrhagic Fever (Sabia virus)

Bunyaviruses: Crimean-Congo Hemorrhagic Fever (CCHF), Rift Valley Fever (RVF)

Flaviviruses: Dengue, Yellow Fever, Omsk Hemorrhagic Fever, and Kyasanur Forest disease [9]

That being said, there are other potential viruses that this data search have identified such as rabies, which is a highly infective agent that if introduced into the food chain, although perhaps, not infective would certainly be likely to cause panic.

Others such as adenovirus which has caused huge outbreaks in susceptible military populations could conceivably be more of a disabling virus that also affected populations [5], while HantaVirus has been associated with recent outbreaks in the United States [7].

Therefore, it is hard to discount completely that in some form that most of the viruses on this list could at least create fear and panic in populations, simply by their introduction – we only need to look at the recent shortage of influenza vaccine to see that populations may not behave rationally in regards to risk when dealing with infectious diseases. Therefore, such a list may at least remind us that there are other viral agents that potentially cause disease and/or terror in populations as well as those commonly known groups.

References

1. The Association of State and Territorial Health Officials (ASTHO). Public Health Preparedness: A Progress Report – First Six Months (ATAIP Indicators Project) (2003)
2. Büchen-Osmond C. Taxonomy and Classification of Viruses. In: Manual of Clinical Microbiology, 8th ed, Vol 2, p. 1217-1226, ASM Press, Washington DC (2003)
3. Diaz, Rumm et. al. National Advisory Committee on Children and Terrorism – Report to the Secretary of DHHS (2003)
4. DiGiacome, R.A, Kremer, J.M. and Shah, D.M. Fish oil dietary supplementation is patients with Raynaud's phenomenon: A double-blind, controlled, prospective study, American Journal of Medicine, 8, 1989, 158-164.
5. Frist B. When Every Moment Counts – What You Need to Know About Bio-terrorism from the Senate's Only Doctor, Rowman and Littlefield (2002)
6. Geissler, E. (Ed.), Biological and toxin weapons today, Oxford, UK: SIPRI (1986)
7. Gray GC, Callahan JD, Hawksworth AK, Fisher CA, and Gaydos JC. Respiratory diseases among U.S. military personnel: countering emerging threats. Emerging Infectious Disease, Vol 5(3): 379-87 (1999)
8. Gursky EA, Drafted to Fight Terror, U.S. Public Health on the Front Lines of Biological Defense, ANSER (2004)
9. Lane SP, Beugelsdijk T, and Patel CK. FirePower in the Lab – Automation in the Fight Against Infectious Diseases and Bioterrorism, John Henry Press, DC (1999)
10. Swanson, DR, Smalheiser NR, & Bookstein A. Information discovery from complementary literatures: categorizing viruses as potential weapons. JASIST 52(10): 797-812 (2001)
11. Swanson, DR., Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 30(1), 7-18 (1986)
12. Swanson, DR., Undiscovered public knowledge. Libr. Q. 56(2), pp. 103-118 (1986)
13. Web site updated regularly by the Agency for Health Care Research and Quality, US DHHS on bioterrorism and emerging infectious disease agents accessible at: <http://www.bioterrorism.uab.edu/EIPBA/vhf/summary.html>