Information Discovery from Complementary Literatures: Categorizing Viruses as Potential Weapons

Don R. Swanson
University of Chicago, 1010 E. 59 St., Chicago, IL 60637. E-mail: d-swanson@uchicago.edu

Neil R. Smalheiser
Psychiatric Institute MC 912, University of Illinois/Chicago, 1601 W. Taylor St., Room 285, Chicago, IL 60612. E-mail: smalheiser@psych.uic.edu

A. Bookstein
University of Chicago, 1010 E. 59 St., Chicago, IL 60637. E-mail: a-bookstein@uchicago.edu

Using novel informatics techniques to process the output of Medline searches, we have generated a list of viruses that may have the potential for development as weapons. 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. This article stresses methods that are more generally relevant to information science. Initial Medline searches identified two kinds of virus literatures—the first concerning the genetic aspects of virulence, and the second concerning the transmission of viral diseases. Both literatures taken together are of central importance in identifying research relevant to the development of biological weapons. Yet, the two literatures had very few articles in common. We downloaded the Medline records for each of the two literatures and used a computer to extract all virus terms common to both. The fact that the resulting virus list includes most of an earlier independently published list of viruses considered by military experts to have the highest threat as potential biological weapons served as a test of the method; the test outcome showed a high degree of statistical significance, thus supporting an inference that the new viruses on the list share certain important characteristics with viruses of known biological warfare interest.

1. Introduction

We have previously conducted studies of complementary noninteractive literatures that can lead to medical discoveries and predictions. Two scientific literatures or sets of articles are complementary if together they suggest new information not apparent in either of the sets considered separately. If, in addition, the two sets are noninteractive in that they do not cite one another and have not been cocited, then the discovered information may be altogether novel. These studies have been documented in the information science and general biomedical literature (Swanson, 1986a, 1986b, 1987, 1988, 1989a, 1989b, 1990a, 1990b, 1990c, 1991, 1993), in the artificial intelligence literature (Swanson & Smalheiser, 1997), and in the neuroscience literature (Smalheiser & Swanson, 1994, 1996a, 1996b, 1998a, 1998b). Medical predictions in three of the earlier studies (Swanson, 1986a, 1988, 1990a) have been corroborated independently by medical researchers. These corroborations have been discussed and cited in (Swanson, 1993; Swanson & Smalheiser, 1999).

The purpose of this project is to demonstrate how the above techniques of analyzing complementary literatures might be applied to problems of defense against biological weapons (*bw*), whether used in the context of warfare or terrorism. The particular problem we selected is that of identifying viruses that might have the potential for development as weapons. This article is based solely on the open-source scientific literature, and is oriented on informatics techniques rather than the virological aspects of the findings. Accordingly, it is written primarily for information scientists and, therefore, focuses more on form than on substantive content of the virus literature (Bates, 1999). At the same time, it is also intended as a guide to the extensive and unmanageably large literature on viruses, as represented by more than a half million records in the Medline database alone. There is extensive open-source literature on biological warfare; we cite here only a few monographs that can serve as a general orientation (Carus, 1991; Geissler, 1986; Roberts, 1993).

The second Review Conference of the Biological Weapons Convention of 1972 took place in 1986. In preparation for this conference, and under sponsorship of the Stockholm

International Peace Research Institute, Erhard Geissler assembled and edited a monograph on biological weapons (Geissler, 1986). From numerous government documents and scientific papers Geissler compiled a list of 21 viruses (in addition to bacteria, toxins, and fungi) known to be regarded by the military as potential biological warfare agents. Geissler identified and summarized 13 criteria underlying these choices (Geissler, 1986, p. 21), quoted below:

> To be an effective agent for use in either a strategic or a tactical role, a biological warfare agent should meet specific criteria (Anderson & King, 1983, p. 3; USA, 1964, p. 25). These criteria can be summarized as follows:
>
> 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. (Geissler, 1986, p. 21)

Geissler states that the agents in his list must meet "at least some" of the above criteria (1986, p. 22).

In response to the Antiterrorism and Effective Death Penalty Act of 1996, the Centers for Disease Control (CDC) amended earlier requirements for the packaging, labeling, and transport of select hazardous infectious agents, and published a comprehensive list of 40 select agents, including 17 viruses (U.S. Code of Federal Regulations, 1999; U.S. Senate, 1998, pp. 22–23). All but four of these viruses are also on the Geissler list. Table 1 provides a list of the 25 virus names on the combined Geissler and CDC lists. For our purposes the CDC contribution serves to give the Geissler list more currency, weight, authority, and credibility.

TABLE 1. Combined Geissler (G) and CDC (C) virus lists.

G	–	Chikungunya Virus
G	C	Crimean-Congo hemorrhagic fever virus
G	–	Dengue Virus
G	C	Eastern encephalitis virus
G	C	Ebola Virus
–	C	Equine Morbillivirus (Hendra virus)
–	C	Flexal virus
–	C	Guanarito virus
G	C	Hantaan (Korean hemorrhagic fever) virus
G	–	Hepatitis A virus
G	–	Influenza virus
G	–	Japanese encephalitis virus
G	C	Junin (Argentine hemorrhagic fever) virus
G	C	Lassa fever Virus
G	–	Lymphocytic Choriomeningitis Virus
G	C	Machupo (Bolivian hemorrhagic fever) virus
G	C	Marburg Virus
G	C	Rift Valley Fever Virus
–	C	Sabio virus
G	–	St. Louis encephalitis virus
–	C	Tick-borne encephalitis viruses
G	–	Russian spring–summer encephalitis virus
G	C	Variola Virus
G	C	Venezuelan encephalitis virus
G	–	Western encephalitis virus
G	C	Yellow Fever Virus

2. Basic Aims and Assumptions

Our objective is to extend the Geissler-CDC (G) list to other viruses that might have similar characteristics. The process we use to extend the list will be tested in part by its ability to yield viruses already on the list. Our basic assumption is that non-G viruses turned up by the same process that yields G-viruses are worth examining for potential *bw* interest—that is, as prospective candidates for extending the G-list. We assume further that the greater the number of G-virus names relative to non-G virus names on any search output, the greater should be our interest in the non-G viruses in the list. Our aim then is to develop lists for which the number of G-terms exceeds what would be expected by pure chance.

3. Hypotheses to be Tested

Most simply, a virus to be considered as a potential weapon must be dangerous, transmissible, and deliverable. This highly condensed statement can be inferred from the 13 criteria summarized by Geissler and quoted earlier (especially from criteria numbered 1, 10, 11, and 12), but disregards many other criteria in his summary that may be important. We propose to identify interesting viruses by processing the output of Medline searches, and the condensed statement provides a reasonable point of departure for constructing appropriate search strategies. There is, however, at least one more basic criterion that should not be disregarded and is well adapted to a Medline search. Viruses not presently known to be dangerous or readily transmissible might become so through genetic alteration. The hy-

potheses we will test take the form of statements about the research literature on viruses, or, equivalently, statements about database searches that define subsets of, or structures within, the total virus literature. Our decision to focus at least initially on Medline is further discussed in the next section.

We hypothesize, therefore, that viruses that have been researched with respect to the characteristics mentioned above are leading candidates for extending the *G-list*. Specifically, we hypothesize that virus subject headings extracted from sets of Medline records on the following topics would contain more terms for viruses of known *bw* interest than would be expected by chance: (i) the genetic aspects of virulence (*vrlg*); (ii) airborne transmission of viral disease (*air*); (iii) stability of viruses in air or aerosol mixtures (*asb*); and (iv) transmission of viruses by arthropods, including insect vectors (*ive*).

Only certain modes of transmission are of practical interest for weaponization. Airborne transmission by means of aerosolization is recognized as an important means of delivering a biological agent, particularly with respect to weapons of mass destruction. The successful use of airborne transmission requires a virus that is viable and stable in aerosol mixtures. Transmission by mosquitoes or ticks is also a credible threat, but perhaps more in the context of terrorism than warfare.

Because we use the Geissler criteria in part to reconstruct the Geissler list, then test our method by measuring how well it succeeds in producing the list, the issue of apparent circularity should be addressed. Indeed, if there were a unique method for deriving either a Medline search or the virus list based on the stated criteria, there would be nothing to test. In fact, each stage of the process is highly problematic, and no two people attempting to reconstruct the G-list from the stated criteria using literature-search techniques are likely to get the same result. But all results are testable, and so may differ greatly from one another in quality, as measured by how well a search succeeds in producing the G-list as output and at the same time is confined to literature having a strong genetics component.

The specific Medline search strategies based on the topics (i) to (iv) are given in Section 5. Additional hypotheses, based on combinations of (i) to (iv), are described in Section 7.

4. Virus Taxonomy, Virus Literature, and Medline

The sixth report of the International Committee on the Taxonomy of Viruses (ICTV, 1996) describes more than 3,600 species of viruses, classified into 50 families, 9 subfamilies, and 164 genera; a searchable database is maintained on the Web. The committee points out that specialty groups keep track of far more (possibly over 30,000) viruses, virus strains, and subtypes (ICTV, 1996, p. 10).

The Medline database was chosen for several reasons as a point of departure for a search of the virus literature. First, indexing is based on a well-developed hierarchical Medical Subject Heading (MeSH) system that facilitates mapping the language of the searcher to the language of the authors of journal articles (NLM, 1999). Second, the headings are applied in depth by skilled indexers who read the entire article in an attempt to provide access to all important concepts. Third, the search command language permits an *exploded* search of all branches beginning at the top of any MeSH tree or subtree, a capability that is important for viruses because there are hundreds of specific virus headings that would be impractical to search individually. The same is true for the category of virus diseases, which occupies a section of the MeSH tree (C2) separate and distinct from the section occupied by viruses (B4).

However, only a small fraction of the total number of identified viruses are assigned Medical Subject Headings. Inspection of the printed MeSH classification schedules does not immediately reveal how many viruses are included. Accordingly, the MeSH-tree schedule for section B4 (viruses) was downloaded from the PubMed site (NLM, 1999) in April 2000, and sorted alphabetically on virus name, removing duplicate names. According to our count, the MeSH scheme includes subject headings for 43 (of the 50) virus families, 70 genera, but only about 230 of the 3,600 individual species identified by the ICTV. Restricting our attention to viruses that can inhabit vertebrates (that is, omitting viruses found only in algae, bacteria, fungi, invertebrates, mycoplasma, plants, and protozoa), as identified in (ICTV, 1996, pp. 26–27), leads to a count of 207 virus terms. One further restriction is imposed, and will be discussed at a later point. Our final value for the size T of the total population that we shall consider is 173. The MeSH hierarchical structure largely reflects the ICTV taxonomic schedules, but there are appreciable differences.

A search of Medline for all articles indexed with virus headings yields close to 320,000 records; about the same number of records are indexed with a virus disease heading (as of May 2000). The intersection of these two sets contains about 120,000 records, and the union about 520,000. The immense size of this literature suggests that it is both fragmented (into numerous subspecialties) and unmanageable.

5. Medline Searches Relevant to Virulence, Genetics, and Virus Disease Transmission

We took the four properties of viruses listed in Section 3 as the basis for developing four Medline searches to be tested as hypotheses. Our aim is to identify which viruses have been investigated with respect to each of these four properties. Before further discussing this general aim, or how the searches relate to one another, we describe each of them individually.

The following four searches were conducted in the Ovid-Medline database covering the period from 1966 to about August 2000. The searches were conducted over a several-week period; slight variations in identically formed search

sets reflect Medline updates, thus introducing small but negligible perturbations into our results.

The first search creates the *virulence-genetic set* (*vrlg*):

1. exp viruses/ge (94910)
2. exp viruses/py (19495)
3. exp virus diseases/mo (6364)
4. 1 and 2 and virulence.sh. (1445)
5. (2 or 3) and virulence/ge (517)
6. 1 and 3 (188)
7. 4 or 5 or 6 (1628) [downloaded to a file named *vrlg*]

The aim of this search is to create a relatively small final set of records on both the genetics and pathogenicity of viruses, where the concept of pathogenicity is supplemented by two related terms—virulence, and disease mortality. Genetics, pathogenicity, and mortality are represented by subheadings (ge, py, and mo). The first three sets are the major building blocks of the search. Each is an *exploded* (exp) subject heading, as defined previously. In sets 4 and 5, we chose to overspecify the concepts of virulence and pathogenicity by forming intersections to make the final set as small and precise as reasonable. Each of the sets 4, 5, and 6 represents one way of getting at the main intent of the search, so the final result was formed by the union of these sets.

The purpose of the next search is to find records on the survivability, viability, or stability of viruses in air or in aerosol mixtures (*asb*):

1. exp viruses/ (323319)
2. air microbiology.sh. or airborne.tw. or aerosols.sh. or aerosol$.tw. (27078)
3. 1 and 2 (1212)
4. (surviv$ or stab$ or viab$).tw. (499076)
5. freeze drying.sh,tw. or (freeze adj dried).tw. (8761)
6. (powder$ or lyophiliz$ or dry).tw. (42127)
7. 3 and (4 or 5 or 6) (147) [downloaded to file *asb*]

There are no subject headings that closely represent the concept of airborne, though the headings *air microbiology* and *aerosols* are useful. Accordingly, these two headings are supplemented with the text (title or abstract) words *airborne* and *aerosol*—as indicated in set 2 above. Set 3, the intersection of sets 1 and 2, is intended to be about airborne viruses. Set 4 is constructed from the stems of three text words that one would expect to encounter in articles that are about the stability, viability, or survival of viruses (in air or aerosol mixtures). Sets 5 and 6 take account of the fact that viruses in a dry or powdered suspension are more stable than viruses in a wet suspension (e.g., aerosol droplets). Set 7 thus contains records on airborne viruses that also mention one or more of these various terms in sets 4, 5, and 6 related to stability. The final result is a local file named *asb* (air stability).

The third search is intended to find records on air or aerosol transmission (*air*) of viral diseases:

1. exp viruses/ (323319)
2. exp virus diseases/ (323308)
3. air microbiology.sh. or airborne.tw. or aerosols.sh. or aerosol$.tw. (27078)
4. 1 and 2 and 3 (823) [downloaded to local file *air*]

Use of the subheading tm (transmission) attached to the virus diseases category (set 2) might seem a reasonable way to narrow down this search. However, the tm subheading did not appear to have been applied consistently enough for our purposes, and so was omitted from the search in the expectation that most records mentioning all three concepts—a virus, a virus disease, and airborne (etc.) probably refer to airborne transmission of the disease.

The fourth search finds records on transmission of viral diseases by insect vectors (*ive*):

1. exp viruses/ (323697)
2. exp virus diseases/ (323671)
3. exp arachnida/ or insect vectors.sh. (23650)
4. 1 and 2 and 3 (1395) [downloaded to local file *ive*]

The subject heading *insect vectors* is useful, but it is apparently limited, as its name would imply, to the Insecta class of Arthropoda—most often mosquitoes. The importance of ticks and mites for viral disease transmission is reflected in set 3. (Arthropod vectors and arachnid vectors are also available headings, but seemed to be less consistently applied than insect vectors.) For the same reason as in the preceding search on air transmission, the subheading tm was not specified for set 2, the *virus diseases* search.

The above four searches forming the sets *vrlg*, *asb*, *air*, and *ive* are not unique, nor can it be proved that they are optimal. We claim only that these searches can be considered as plausible hypotheses that are testable (Harter, 1984)—the hypotheses being that the searches yield statistically significant results when tested as described in Section 8.

We return to the question raised earlier of how these four sets are, or should be, related to one another. To seek articles that are about all four sets is too specific, for the intersection "*vrlg* AND *asb* AND *air* AND *ive*" yields nothing at all. Because being in *vrlg* is a *sine qua non* for being of *bw* value, it is reasonable to ask for all articles that are in both the virulence-genetics (*vrlg*) set, and any one of the three transmission-related sets (*asb*, *air*, or *ive*). The four searches were conducted as a single on-line search, and these three intersections were formed, but the retrieved sets turned out to be very small (0, 1, and 13 records, respectively). That is, there are almost no records in the Medline database on both of the two crucial virus properties of virulence-genetics and disease transmission as represented in our four Medline searches. Accordingly, we embarked on a new approach to database searching that we have successfully applied to other problems.

6. A Basic Limitation of Conventional Database Searching, and a Revised Statement of the Problem

Conventional database searching is based on forming and combining sets of bibliographic records, and so is expressed in terms of the desired properties of those records. That is, even though we are actually seeking viruses having certain properties, to conduct a database search we must express virus properties as properties of the records that are presumed to contain the desired information. The most notable result of doing so in this case, as described above, was the virtual absence of records that met our search specifications.

However, if we restate the problem directly in terms of finding all viruses (rather than all records) that have been investigated in terms of both *vrlg* and any of the three transmission related terms, we are led to a new kind of solution that does not depend on finding any records or articles at all that individually fulfill the search criteria. To see how this might occur, suppose, for example, that, for some virus unknown to the searcher, the properties corresponding to *vrlg* and, say, *airborne transmission* have been investigated in two separate articles—one article on *vrlg* and one on *airborne transmission*—but that no single article on that virus exists that reports work on both sets of properties. Thus, a conventional search (which cannot include the name of the virus, because it is presumed unknown) based on forming an intersection of properties such as *vrlg* AND *airborne transmission,* would yield no articles at all on that particular virus.

To generalize, if we find that a set such as *vrlg* does not intersect a set of records on *airborne transmission,* it is directly relevant to our problem to ask which virus terms are common to the two sets. There is no practical way to answer this question using conventional database search techniques that produce a null intersection, but it is the general type of question that can be answered by our techniques of analyzing complementary noninteractive literatures with the aid of software that we have named Arrowsmith. Arrowsmith can extend the power of database searching by automatically processing the downloaded results of a search. It can, for example, find all virus terms (or other subject headings, or words and phrases) that are common to two downloaded sets of records, and then organize and display this information in a way that helps the user see new connections of scientific interest.

7. Outline of Arrowsmith Procedure for the Virus Problem

Arrowsmith always takes as its input a single pair of sets of document records (A,C), and produces as its output a list of terms common to both sets, which we refer to as a B-list. For our purposes, the set that we denote as C is the downloaded search output *vrlg.* The sets *asb, air,* and *ive* are taken, in turn, as the set A. That is, we seek viruses that are both virulent and transmissible, and because the transmission category is divided into three parts (*airborne, stability in air,* and *insect vectors*), we therefore consider three A–C pairs: *asb-vrlg, air-vrlg,* and *ive-vrlg.* These pairs become three separate inputs to Arrowsmith, and Arrowsmith produces the three corresponding B-lists of virus terms.

A comment on notation: *vrlg, asb, air,* and *ive* are sets of Medline records downloaded from the four searches described in Section 5. From here on, the three hyphenated pairs (*asb-vrlg, air-vrlg, ive-vrlg*) will be used to denote the three B-lists from Arrowsmith; they are each lists of virus names. Because *vrlg* is common to all three, the first member of each hyphenated pair will be used as an abbreviation for the pair wherever the context makes clear that this second meaning is intended rather than the above Medline sets.

Each B-list is a list of subject headings that occur in both A and C, even if A and C have no records in common. Virus terms extracted or selected from the B-list represent a list of viruses that have been investigated with respect to both A and C, which responds exactly to the most salient goal of the search process, as reformulated above. For example, in the case of the *asb-vrlg* B-list, we would like to know which viruses have been investigated with respect to both (a) the genetic aspects of virulence, and (b) stability in aerosols—particularly if such investigations have been conducted separately and independently, for, in that case, the fact that both investigations have been conducted may have gone unnoticed. In effect, our aim in this project is to categorize viruses in terms of the state of their research literatures that may have previously unnoticed relevance to biological weapons. We stress the word "may" in the two preceding sentences because we are never in a position to know with certainty that no one has noticed particular implicit connections in the literature. We focus only on the state of the literature and on whether it tends to facilitate or hinder seeing new connections. In general, it is more difficult to find implicit connections between text passages in two different articles than to find connections that are wholly contained within one article, and which, therefore, are usually explicit.

Combining the Three B-Lists into Virus List 1 (VL1)

From each of the three B-lists identified above, virus headings were selected automatically with the help of the downloaded MeSH-B4 tree structure, and processed into a list of virus species, as described in the first part of this article. (MeSH terms for virus genera are analyzed separately in Appendix I.) The automatic selection of virus terms represents a step toward further automation of Arrowsmith, insofar as desired categories of terms in the B-list can be identified in advance from a controlled vocabulary such as MeSH. Otherwise, as is generally the case for title-word processing, the B-list is edited by the user.

Some of the selected viruses will be on all three B-lists, some on two of the three, and some on just one list. It is

plausible to assume that our interest in a virus as a *bw* prospect may depend on just which lists it is on. Accordingly a combined list called VL1 is formed (Table 2), in which each virus term is accompanied by the code *asb, air,* and/or *ive* (which, within VL1, is a shorthand notation for *asb-vrlg, air-vrlg,* and *ive-vrlg,* resp.). Any single code, any two codes, or all three codes may be present.

VL1 consists of raw data that must be evaluated to determine whether the intended term relationships within each search were met. But this further work of search evaluation requires human judgment, which is difficult to reduce to a formal description that is reproducible and testable. VL1, on the other hand, was produced by an automatic, reproducible process (given the initial Medline searches described earlier), and can be objectively tested, as will be shown. In effect, what is tested are the Medline searches that produced the four files *vrlg, asb, air,* and *ive,* as well as the Arrowsmith procedure for processing the search results into the three virus B-lists *asb-vrlg, air-vrlg,* and *ive-vrlg.*

Because VL1 contains unevaluated results of the four Medline searches, it is only the first step toward a ranked list useful for the purpose of assessing potential *bw* defense interest. Until the list is evaluated and edited, any of the listed terms in VL1 may be a spurious or accidental search outcome. But this first step, production of the raw search data, is at least reproducible and is, therefore, testable for statistical significance. Later we evaluate each search, eliminate spurious results, and revise our estimates of statistical significance accordingly.

8. Statistical Tests of the Medline and Arrowsmith Output for VL1

Our objective here is first to test the efficacy of our initial Medline searches in producing the 25-term Geissler-CDC list, and then compare this to the success of the Arrowsmith approach to processing the output of the Medline searches. To test each Medline/Arrowsmith search separately, we computed the probability that the observed number of G-terms, or more, could be attributed to chance.

To compute these probabilities, we first need to identify the *universe* of viruses from which all virus terms are selected. If the size of this universe is T, then the elementary probability, $1/T$, of randomly selecting one virus from this universe is the crux of a random model. As we noted in Section 1, the ICTV (1996, p. 10) mentioned 30,000 viruses, virus strains, and subtypes as a possible universe, but their own taxonomy spans only about 3,600. For purposes of designing a valid test, it is important not to use an inflated universe that includes viruses that for one reason or another could not or would not have been considered for the Geissler list. An inflated T would lead to an abnormally low value for the probability $1/T$ and, hence, bias the test toward a favorable result. The universe is difficult to define, but our approach here is to make it as restrictive or deflated as reasonable to conduct a stringent test.

A study of the Geissler list shows that essentially all of its entries have corresponding MeSH terms at the taxonomic level of the species, and all viruses listed can infect vertebrates. In Section 4 we noted that 207 virus terms qualify. A few adjustments were needed to achieve a match between the Geissler-CDC list and its MeSH equivalents; these are discussed in Appendix I, which includes also an analysis of the need to select certain viruses at the genus level rather than the species. These details may be of more value to someone trying to replicate the search results than to the more casual reader who might prefer to skip Appendix I.

One more deflationary factor seems required. Some of the virus headings or virus names were not available at the time the Geissler list was created, and so the value of T should be reduced by that amount. With the help of Medline searches restricted to pre-1986, and with help from dates given in the published alphabetic MeSH manual, we determined that 34 terms should be omitted, yielding a final value of $T = 173$. (These 34 terms were not omitted in the search and download process so as not to exclude the more recent literature; this reduction of T affects only the ensuing probability calculations, and then only by altering the total count).

A further comment is in order with respect to the probabilistic assumptions that underlie statistical testing. In a Medline search, the a priori probability of selecting a virus term is not the same for all terms. The total number of Medline records for any given virus heading is correlated with the chance of its being selected. The possibility must be considered that the G-terms might be viruses that are the most extensively researched and have the largest literatures, and so our Medline search process might tend to capture a disproportionately large number of them. However, this possibility was checked and can reasonably be dismissed. The average number of records in the Medline database for the G-terms is 806, while the average over the rest of the 207 terms (i.e., before eliminating the 34 unavailable terms) is 1027. Thus there seems to be no tendency for the G-viruses to have the largest literatures.

VL1 reflects the combined effects of Medline and Arrowsmith, but the Medline searches that form the input to Arrowsmith can be tested independently of Arrowsmith by transforming each search output (*asb, air, ive, vrlg*) into a virus list. It was found that only one of the four files (*ive*) contained a clearly statistically significant number of G-virus terms, with a second (*air*) being suggestive: (*asb:* $p = 0.098$; *air:* $p = 0.055$; *ive:* $p = 0.000040$; and *vrlg:* $p = 0.37$). Values of p are calculated using a hypergeometric distribution.

After Arrowsmith processing, however, the picture changes. The 90-term VL1 list (Table 2) contains 20 G-terms. Again, calculated from a hypergeometric distribution, the probability that 20 or more could occur by chance is 0.0020. Each of the three B-lists (the Arrowsmith output), considered independently, is also statistically significant in the number of G-terms that occur: *asb-vrlg:* $p = 0.044$; *air-vrlg:* $p = 0.015$; and *ive-vrlg:* $p = 0.0000007$. Ar-

TABLE 2. Virus list VL1 (raw data).

–	0–	–––	–––	air	Adenoviruses Canine
–	––	–––	–––	air	Adenoviruses Human
–	0–	ive	asb	air	African Horse Sickness Virus
–	––	–––	ive	air	African Swine Fever Virus
–	––	ive	asb	air	Aphthovirus
–	1x	–––	–––	ive	Bluetongue Virus
G	––	–––	–––	ive	Chikungunya Virus
–	––	–––	–––	air	Coxsackieviruses A
–	––	–––	–––	air	Coxsackieviruses B
G	––	–––	ive	aix	Dengue Virus
–	––	ive	asx	air	Diarrhea Virus Bovine Viral
–	––	–––	–––	air	Distemper Virus Canine
G	––	ivx	asx	air	Ebola Virus
–	2x	–––	–––	air	Ectromelia Virus
G	––	–––	–––	ive	Encephalitis Virus Eastern Equine
G	––	ive	asb	air	Encephalitis Virus Japanese
–	0–	–––	–––	ive	Encephalitis Virus Murray Valley
G	––	–––	ive	air	Encephalitis Virus St. Louis
G	––	ive	asb	air	Encephalitis Virus Venezuelan Equine
G	––	–––	–––	ive	Encephalitis Virus Western Equine
G	––	–––	ive	asb	Encephalitis Viruses Tick-Borne
–	––	ivx	asb	aix	Encephalomyocarditis Virus
–	0–	–––	–––	ive	Ephemeral Fever Virus Bovine
–	0–	–––	ive	air	Fowlpox Virus
–	––	–––	–––	asx	Gastroenteritis Virus Murine
–	––	–––	ive	aix	Gastroenteritis Virus Porcine Transmissible
–	––	–––	ivx	aix	HIV
–	––	ivx	asx	aix	HIV-1
–	––	–––	–––	ivx	HTLV-BLV Viruses
G	––	ive	asb	air	Hantavirus (or Hantaan virus)
G	––	–––	–––	ive	Hemorrhagic Fever Virus Crimean-Congo
–	––	–––	–––	ive	Hepatitis Agents GB
–	––	–––	ive	aix	Hepatitis B Virus
–	––	–––	–––	ivx	Hepatitis C-Like Viruses
G	––	–––	ive	air	Hepatovirus (Hepatitis A virus)
–	––	ivx	asb	air	Herpesvirus 1 Bovine
–	––	–––	–––	aix	Herpesvirus 1 Cercopithecine
–	1x	–––	ive	air	Herpesvirus 1 Equid
–	0–	–––	asb	air	Herpesvirus 1 Gallid
–	––	–––	–––	air	Herpesvirus 1 Human
–	––	ive	asb	air	Herpesvirus 1 Suid
–	––	–––	–––	air	Herpesvirus 2 Gallid
–	––	–––	–––	air	Herpesvirus 3 Human
–	––	–––	–––	ive	Herpesvirus 4 Human
–	––	–––	–––	air	Herpesvirus 6 Human
–	0–	–––	asx	air	Hog Cholera Virus
–	1x	–––	–––	ive	Infectious Anemia Virus Equine
–	0–	–––	–––	air	Infectious Bronchitis Virus Avian
–	––	–––	–––	air	Infectious Bursal Disease Virus
–	1x	–––	asx	air	Infectious Peritonitis Virus Feline
–	––	–––	–––	air	Influenza A Virus Avian
G	––	ive	asb	air	Influenza A Virus Human
–	––	–––	–––	air	Influenza A Virus Porcine
–	––	–––	–––	air	Influenza B Virus
–	1x	–––	–––	ive	Kemerovo Virus
G	––	ive	asb	air	Lassa Virus
–	––	–––	–––	ive	Leukemia Virus Bovine
–	––	–––	–––	ivx	Leukemia Viruses Murine
–	––	–––	–––	air	Leukosis Virus Avian
G	––	–––	ive	air	Lymphocytic Choriomeningitis Virus
G	––	ive	asb	air	Marburg Virus
–	––	–––	asb	air	Measles Virus
–	––	–––	–––	asb	Moloney Leukemia Virus
–	––	–––	–––	ivx	Monkeypox Virus
–	––	–––	asx	air	Mumps Virus
–	––	–––	–––	ive	Myxoma Virus

TABLE 2. (continued)

–	– –	ive	asb	air	Newcastle Disease Virus
–	– –	– – –	asx	aix	Papillomavirus Human
–	– –	– – –	asx	air	Parainfluenza Virus 1 Human
–	– –	ive	asb	air	Parainfluenza Virus 3 Human
–	– –	– – –	– – –	asb	Polyomavirus macacae
–	– –	– – –	ive	air	Rabies Virus
–	– –	– – –	ivx	air	Respiratory Syncytial Virus Human
–	– –	ive	asx	air	Respiratory Syncytial Viruses
–	– –	ive	asb	air	Rhinovirus
G	– –	– – –	ive	air	Rift Valley Fever Virus
–	– –	ivx	asb	air	Rinderpest Virus
–	– –	– – –	– – –	ive	Ross River Virus
–	– –	– – –	asb	air	Rotavirus
–	– –	– – –	ivx	aix	SIV
–	– –	– – –	– – –	ivx	Sarcoma Viruses Simian
–	– –	ive	asb	aix	Semliki Forest Virus
–	– –	– – –	– – –	ive	Sindbis Virus
–	– –	– – –	asx	aix	Swine Infertility and Respiratory Syndrome Virus
G	– –	– – –	ive	air	Tacaribe Complex Viruses
–	– –	ivx	asb	air	Vaccinia Virus
G	– –	– – –	ive	air	Variola Virus
–	– –	ive	asb	aix	Vesicular Stomatitis-Indiana Virus
–	– –	– – –	ive	air	West Nile Virus
G	– –	ive	asb	aix	Yellow Fever Virus

G in col 1 denotes Geissler list, CDC list, or both. Each virus entry on list VL1 is coded to represent the Medline search categories and Arrowsmith output categories in which it fell. The codes are: asb = *asb-vrlg;* air = *air-vrlg;* ive = *ive-vrlg,* as defined in the text. asx, aix, and ivx denote evaluated searches that were found spurious. The numbers 0, 1, or 2 in col. 2 mark all virus terms with fewer than three records that intersect the set of all records on human pathogenicity.

Terms with only one or two intersecting records were evaluated by examining abstracts; x denotes negative evaluation outcome (i.e., records do not show that the virus is pathogenic in humans). This point is further discussed in Appendix 2 and Section 9. (x in all cases denotes a search that did not have the intended outcome, as judged from the abstracts.)

rowsmith appears to have a strong and favorable effect on the proportion of G-terms within each output list. Consistent with the Medline output, *ive-vrlg* is most clearly significant, with *air-vrlg* following. More detailed data are given in Table 3.

We next analyze how combinations of variables contribute to the statistical significance of our results. Each virus term falls into one of eight categories (labeled c1 through c8), according to which, if any, combination of the three lists it appears in. Also, each virus falls uniquely into one of two categories, G or not-G, according as it is or is not on the Geissler-CDC list. Table 4A (raw data) presents a count of the number of terms in each of the 16 categories, displayed as a two-row by eight-column table.

The numbers of G-virus names that appear in each of the 16 categories of VL1 (Table 4A) were used as the basis for more detailed testing of the procedure that produced the VL1 list. A chi-squared test of the complete table shows that it is statistically significant [$p < 0.001$, 7 degrees of freedom (*df*)]. However, our main interest lies in determining the significance of the individual categories. Because category c8 corresponds to membership in none of the lists, as indicated in Table 4A, we tested each of the other 7 (c1–c7) against c8. Everitt (1977, pp. 44–46) shows a "reasonable" (but conservative) way to do this, with attribution to M.N. Brunden, by partitioning the two by eight 7-*df* table into seven 2×2 1-*df* tables (corresponding to c1:c8, c2:c8, etc.), then testing each one at a significance level much more stringent than the initially chosen level. If the overall target level is 0.05 for 7-*df,* then each 2×2 test should be conducted at a significance level of 0.05 divided by two-times $(df - 1) = 0.004$ with the new $df = 1$. From Table 4A, we see that only categories c1 (viruses on all three lists) and c5 (viruses on the 2 lists *air-vrlg* and *ive-vrlg*) are significant. This result is tentative because of the possible influence of spurious search outcomes, which we discuss further.

9. Refining VL1

We next assessed some of the search outcome. Our particular concern was virus terms for which there were only one or two postings in the categories *asb, air,* and *ive,* for these were the cases most prone to change on evaluation because a single record could determine whether a virus was included or excluded from the corresponding category. The evaluation outcome was used to identify aberrations in which the results of a search did not satisfy what was intended by the search. The results of each evaluation in which a search outcome was judged to be spurious is coded

TABLE 3. Effect of Arrowsmith on statistical significance.

1 List	2 Total	3 Non-G	4 G-terms	5 Expected G	6 %G-terms	7 *p* (hypergeom)
all	173	148	25	25.0	14%	1.000 n.s.
vrlg	137	116	21	19.8	15%	0.367 n.s.
asb	41	32	9	5.9	22%	0.098 n.s.
asb-vrlg	36	27	9	5.2	25%	0.044
ti asb-vrlg	33	25*	8	4.8	24%	0.07 n.s.
air	75	60	15	10.8	20%	0.055 n.s.
air-vrlg	66	51	15	9.5	23%	0.015
ti air-vrlg	58	48*	10	8.4	17%	0.30 n.s.
ive	73	53	20	10.6	27%	0.000040
ive-vrlg	61	41	20	8.8	33%	0.000001

cols. 2, 3, 4, 5 show numbers of virus terms.

Rows labelled vrlg, asb, air, and ive are the initial Medline search results that are used (in pairs) as input to Arrowsmith.

Rows asb-vrlg, air-vrlg, and ive-vrlg show Arrowsmith output.

The probability (col. 7) of selecting as many G-terms as shown in col. 4, or more, by a purely random process is calculated using the cumulative hypergeometric distribution; the significance criterion is taken as 0.05.

The expected values of G for a given Total (col. 2) are in col. 5.

* ti in left margin indicates title processing (See Appendix III).

* Number of Non-G for titles is only approximate.

TABLE 4. Partitioning a 2 × 8 contingency table (7-*df*): testing for independence of G vs. non-G.

1, 0 denote presence or absence of asb-vrlg, air-vrlg, or ive-vrlg for each virus entry—shown in that order as a set of three in the tabulation below. The numbers in the table show the number of virus terms for each triplet. G refers to the Geissler-CDC list. YCC denotes Yates continuity correction to chi-squared for 1-*df*. Note: Some of the cells in the tables have very few elements; the accuracy of chi-squared tests in such cases is questionable, but a poor approximation in all such cases here would lead to false nonsignificance rather than to false significance. Moreover, using the hypergeometric distribution to calculate *p* (instead of chi-squared tables) does not change the results.

Table A—VL1 (raw data)

	c1 111	c2 110	c3 101	c4 100	c5 011	c6 010	c7 001	c8 000	Total
G	8	0	1	0	7	0	4	5	25
non-G	15	9	0	3	10	17	16	78	148
									173

Overall chi-squared for Table A = 33.6 $p < 0.001$ (7-*df*). Following method of Brunden (Everitt, 1977, p. 44), each of the first seven categories was tested against c8, taking target significance at $0.05/(2(df - 1)) = 0.004$, at 1-*df*. (This formula represents a tradeoff between degrees of freedom and required precision or stringency of the test.) Only the c1 and c5 categories were significant.

c1 v c8: chi-squared = 13.8 or 11.3 with YCC $p < 0.001$ (1-*df*).

c5 v c8: chi-squared = 16.5 or 13.4 with YCC $p < 0.001$ (1-*df*).

Table B—VL2 (search evaluation to refine VL1)

	c1 111	c2 110	c3 101	c4 100	c5 011	c6 010	c7 001	c8 000	Total
G	6	0	2	0	6	1	5	5	25
non-G	6	6	2	3	7	20	13	90	147
									172

Overall chi-squared for Table B = 38.9 $p \ll 0.001$ (7-*df*). Following the same procedure as for Table A, the first and fifth categories remained highly significant (target 0.004):

c1 v c8: chi-squared = 23.1 or 18.5 with YCC $p < 0.001$ (1-*df*).

c5 v c8: chi-squared = 20.9 or 16.7 with YCC $p < 0.001$ (1-*df*).

c3 v c8: chi-squared = 11.7 or 5.9 with YCC $p < 0.02$ (1-*df*) n.s.

c7 v c8: chi-squared = 9.51 or 6.9 with YCC $p < 0.01$ (1-*df*) n.s.

Using the hypergeometric calculation for *p*, all inequalities are the same, except for c3 vs. c8: $p = 0.024$.

for each virus term in VL1 (Table 4A), but without making any deletions. That code consists of replacing the trailing letter of "asb," "air," and/or "ive" with an "x," to give "asx," "aix," or "ivx," respectively, thus showing the specific search category that was judged spurious. Those virus terms for which all three of its transmission codes have an "x" were then deleted, to produce a new list VL2; terms with one or two x-codes were reclassified according to the number of their remaining nonspurious codes.

Table 4B is similar to Table 4A, except that it is based on list VL2 instead of VL1. The new value of the significance probability, *p* (following the x-based deletions), was calculated. Because the x-codes were based on judgment, a bias could, in principle, be introduced into this new test of significance, perhaps, for example, by judging G-virus records more leniently than non-G. It was important to know, however, whether any of the significance vanished after the deletion process, raising the question of whether the significance might have been based solely or very strongly on spurious search outcomes. In fact, significance increased (i.e., *p*-values decreased). We adopted the conservative strategy of assessing results in the light of both VL1 and VL2.

Appendix II describes some further refinements of list VL1. We explore briefly the terms for human pathogenicity in the list VL1. Whether such a criterion should be included in the search specifications is not obvious (one might not want to reject a virus that is highly pathogenic only in animals—either because it might have potential as an agent for destroying farm animals or because it might be altered for use against humans). Accordingly, we did not reject any virus term on the grounds that it was not pathogenic in humans, but we did identify such terms, leaving it up to the user of the list to either keep or reject these terms. We assumed without further evaluation that all virus terms with three or more human–pathogenicity postings probably represented a valid relationship (noting that all G-viruses had four or more such postings). We then marked (in col 2, Tables 2 and 5) all terms with 0, 1, or 2 postings, and appended an "x" if, after examining the abstracts of those records, the relationship appeared to be spurious.

It is plausible to consider viruses that meet more of the 13-item Geissler criteria list as meriting higher rank than those meeting fewer criteria. Accordingly, we rank VL2 according to whether they meet all three criteria, or just two, or just one, as reflected in the following scheme, which will be applied to Table 5, and then go on to discuss the statistical data that do or do not support this decision. The significance probabilities, *p*, are taken from Table 4B, which provides more detailed data. Note that the very stringent condition required above for statistical significance of the multiple comparisons is continued here.

Rank 1: *ive-vrlg* and *air-vrlg* and *asb-vrlg* $p < 0.001$
Rank 2: *ive-vrlg* and *air-vrlg* $p < 0.001$
Rank 3: *ive-vrlg* and *asb-vrlg* (too few terms for reliable *p*-value)
Rank 4: *ive-vrlg* $p = 0.01$ n.s.
(Rank 4 becomes significant ($p < 0.001$) if the five virus terms coded for nonpathogenicity in humans, as assessed in Appendix II, are removed.)

Technically, Rank 4 (category *ive-vrlg* only) is statistically nonsignificant ($p < 0.01$, 1-*df*) because of the stringent criterion $p < 0.004$ being observed, in contrast to Rank 1 and Rank 2 (both $p < 0.001$, 1-*df*), which do meet the 0.004 criterion by a wide margin. This fact might support the rejection of Rank 4 altogether. However, further tests show that it is also true that none of the ranks are statistically distinct from each other when compared directly with each other. Moreover, Rank 4 is worth keeping because any user of the list may choose to discard viruses that are not known to be pathogenic in humans; as noted above, such deletions would make rank 4 significant.

The strongest feature of the data is the dominance of the *ive* category arising from the fact that 19 of 25 members of the G-list in VL2 are associated with insect or arachnid transmission, an effect that is even stronger if the Geissler list is considered separately from the CDC list for then the above result becomes 19 of 21. This finding invites explanation from experts in virology or viral ecology.

The ranking scheme as shown above may be helpful in answering questions about airborne transmission or stability in air, while disregarding insect or tick-borne transmission. Irrespective of statistical significance, it is more conservative and potentially more useful at this point to retain the separate ranks on the grounds of plausibility.

10. A Predictive Model

In the previous sections we described three Arrowsmith-mediated searches that attempted to generate a list of viruses of *bw* threat. To test the efficacy of these methods, we probed how well they retrieved viruses on the G-list, while rejecting other viruses in the Medline database. Some of these latter viruses may well also be of *bw* value, but presumably most are not, and therefore, this group could serve as a contrasting population regarding their bibliographic properties.

Several approaches were used. The first constructed sequences of contingency tables, and used conservative tests of independence. But relying on multiple tests led us to adopt a stringent criterion that may have eliminated promising categories of viruses. To assess in a more detailed manner how appearances of a virus in the various sets influenced the likelihood that the virus appear in the G-list, we modeled the probability of a virus being on the G-list using the well established method of *logistic regression* (Agresti, 1990).

To do this, we use as predictors the following three indicator variables, defined in Table 4A:

asb: Takes the value 1 if the virus appears in the list *asb-vrlg;* else 0.

TABLE 5. Virus list VL2.

Rank 1: all three criteria (ive-vrlg, asb-vrlg, air-vrlg)					$p < 0.004$
G	––	ive	asb	air	Encephalitis Virus Japanese
GC	––	ive	asb	air	Encephalitis Virus Venezuelan Equine
GC	––	ive	asb	air	Hantavirus (Hantaan Virus)
G	––	ive	asb	air	Influenza A Virus Human
GC	––	ive	asb	air	Lassa Virus
GC	––	ive	asb	air	Marburg Virus
––	––	ive	asb	air	Aphthovirus
––	0–	ive	asb	air	African Horse Sickness Virus
––	––	ive	asb	air	Herpesvirus 1 Suid
––	––	ive	asb	air	Newcastle Disease Virus
––	––	ive	asb	air	Parainfluenza Virus 3 Human
––	––	ive	asb	air	Rhinovirus
Rank 2: two criteria (ive-vrlg, air-vrlg)					$p < 0.004$
G–	––	–––	ive	air	Encephalitis Virus St. Louis
G–	––	–––	ive	air	Hepatovirus
GC	––	–––	ive	air	Junin Virus (Tacaribe Complex Viruses)
G–	––	–––	ive	air	Lymphocytic Choriomeningitis Virus
GC	––	–––	ive	air	Rift Valley Fever Virus
GC	––	–––	ive	air	Variola Virus
––	––	–––	ive	air	African Swine Fever Virus
––	––	ive	asx	air	Diarrhea Virus Bovine Viral
––	0–	–––	ive	air	Fowlpox Virus
––	1x	–––	ive	air	Herpesvirus 1 Equid
––	––	–––	ive	air	Rabies Virus
––	––	ive	asx	air	Respiratory Syncytial Viruses
––	––	–––	ive	air	West Nile Virus
Rank 3: two criteria (ive-vrlg, asb-vrlg)					$p < 0.02$ (too few terms for reliable test)
GC	––	–––	ive	asb	Encephalitis Viruses Tick-Borne
GC	––	ive	asb	aix	Yellow Fever Virus
––	––	ive	asb	aix	Semliki Forest Virus
––	––	ive	asb	aix	Vesicular Stomatitis-Indiana Virus
Rank 4: single criterion–––ive-vrlg only					$p < 0.01$ n.s.
G	––	–––	–––	ive	Chikungunya Virus
G	––	–––	ive	aix	Dengue Virus
GC	––	–––	–––	ive	Encephalitis Virus Eastern Equine
G	––	–––	–––	ive	Encephalitis Virus Western Equine
GC	––	–––	–––	ive	Hemorrhagic Fever Virus Crimean-Congo
–	1x	–––	–––	ive	Bluetongue Virus
–	0–	–––	–––	ive	Encephalitis Virus Murray Valley
–	0–	–––	–––	ive	Ephemeral Fever Virus Bovine
–	––	–––	ive	aix	Gastroenteritis Virus Porcine Transmissible
–	––	–––	–––	ive	Hepatitis Agents GB
–	––	–––	ive	aix	Hepatitis B Virus
–	––	–––	–––	ive	Herpesvirus 4 Human
–	1x	–––	–––	ive	Infectious Anemia Virus Equine
–	1x	–––	–––	ive	Kemerovo Virus
–	––	–––	–––	ive	Leukemia Virus Bovine
–	––	–––	–––	ive	Myxoma Virus
–	––	–––	–––	ive	Ross River Virus
–	––	–––	–––	ive	Sindbis Virus

List VL2 was derived from VL1 by by ignoring all "x" terms within fields 3, 4, and 5–––(ivx, asx, and aix), then ranking all terms according to the number of remaining criteria. 0 or 1x in the second field marks the virus as probably nonpathogenic in humans, with either no records, or just one record, that intersects the set of records on human pathogenicity. Rank 4 becomes significant if these virus terms are removed, for they are all in the non-G category.

air: Takes the value 1 if the virus appears in the list *air-vrlg;* else 0.

ive: Takes the value 1 if the virus appears in the list *ive-vrlg;* else 0.

We wish to see how well we can predict a virus's appearance in the G-list, given its appearance in any of the Arrowsmith-constructed lists, as indicated by its values for the indicator predictors. This task is complicated by the possibility that the predictors interact, that is, that the value on one of the predictors influences the predictive strength of the others. Logistic regression allows us to tease out the effect of each predictor, including the possibility of inter-

TABLE 6. Predictive model results for VL2.

	Category	c1	c2	c3	c4	c5	c6	c7	c8	Total
	See Table 4B:	111	110	101	100	011	010	001	000	
G-terms	Observed	6	0	2	0	6	1	5	5	25
	Estimated	5.8	0.5	1.6	0.2	5.5	1.3	6.1	4.1	25
Non-G	Observed	6	6	2	3	7	20	13	90	147
	Estimated	6.2	5.5	2.4	2.8	7.5	19.7	11.9	90.9	147

Comparison of observed values and values predicted by model. The categories denote, respectively: (asb, air, ive), (asb, air), (asb, ive), asb, (air, ive), air, ive, and null.

action. The possibility of deleting (or rejecting) such interactions is a further advantage of using such a statistical model instead of the more standard means of analyzing a complex contingency table.

Specifically, we hypothesize the following model, in which the probability, p, of a virus appearing in the G-list is given by:

$$\log \frac{p}{(1-p)} = \lambda_0 + \lambda_{asb}\mathbf{asb} + \lambda_{air}\mathbf{air} + \lambda_{ive}\mathbf{ive} + \lambda_{asb,air}\mathbf{asb \cdot air} + \lambda_{asb,ive}\mathbf{asb \cdot ive} + \lambda_{air,ive}\mathbf{air \cdot ive} + \lambda_{asb,air,ive}\mathbf{asb \cdot air \cdot ive} \quad (1)$$

In this most general equation, we express the *logistic transform* of p as a linear combination of all the predictors. The logistic transform has the desirable property that all real numbers are legitimate values, in contrast to p itself, which is restricted to values between zero and 1. But also, this transform has been found empirically to be particularly useful in relating probabilities to categorical predictors of the type we have here. The product terms allow us to incorporate the possibility of interactions. For example, the term **asb-air** takes the value 1 only if the virus appears in *both* the *asb* and *air* lists, and thus allows us to correct for the possibility that appearing in the *asb* list influences the impact of a virus's appearing in the *air* list.

Because we have eight cells and eight variables, this equation can always produce a perfect fit. The statistical issue is whether a good fit is possible even if some of the terms are set to zero. In general, we prefer the simplest model that adequately fits the data—that is, we prefer the highest order interaction terms to be zero. To assess this for VL2, we used the *Minitab* statistical package (Minitab, 1998, version 12) to analyze the data, trying a range of models. We found, in fact, that the following simple model, omitting all interactions, fitted the VL2 data very well:

$$\log(p/(1-p)) = -3.09 + 0.24\mathbf{asb} + 0.35\mathbf{air} + 2.43\mathbf{ive} \quad (2)$$

Actually, our analysis finds that only the constant term and **ive** are statistically significant, with values that are about seven and four standard deviations greater than zero, respectively (**asb** and **air** are, respectively, 0.41 and 0.65 standard deviations above zero). We include the other terms for completeness, because of our concern that the lack of significance might reflect the small amount of data available to us.

One can interpret this equation in several ways. Particularly illuminating is to take the exponential of both sides, getting:

$$p/(1-p) = \exp(-3.09)\exp(0.24\mathbf{asb})\exp(0.35\mathbf{air})\exp(2.43\mathbf{ive}) \quad (3)$$

That is, the odds of a virus being on the G-list is given as the product of a constant factor and the factor(s) corresponding to the list(s) in which the virus appears. The constant term, $\exp(-3.09) = 0.05$, gives the odds for the case in which the virus appears in none of the predictor lists, and its small value merely reflects the relative sizes of the G compared to the non-G lists. Each exponential term then indicates the factor by which appearance in the corresponding list increases this odds. These factors are:

asb: 1.28
air: 1.42
ive: 11.40.

The high value for the factor by which **ive** increases the odds reflects its being the only one that is statistically significant. The absence of any interaction term indicates that the factor by which the appearance of a virus in a list increases the odds of its being a G-list virus does not depend on whether the virus appears in any of the other predictor lists.

The following table, comparing the predicted and actual number of viruses on the G-list and not on the G-list in each of five categories (some categories having been merged automatically by Minitab to compensate for small numbers of viruses in some cells), can be used to judge the validity of the model as applied to VL2 (Table 6).

As a formal measure of goodness of fit, the chi-square value is 1.51, on 4 degrees of freedom. Chi-square measures the discrepancy between the data and the values predicted by the model, a large value indicating a model's inadequacy. If the model is valid, the chi-square will be at least

this large 82% of the time, purely as the result of random fluctuation.

11. The Meaning and Robustness of VL2

It is reasonable to ask whether VL2 is "robust" in the sense that other plausible Medline search strategies at the outset would have led to approximately the same list. We explored a number of variations of the initial *vrlg* searches that retained the same broad strategy based on virulence, pathogenicity, and genetics; most variations we explored had only a minor impact on the final list. However, major changes, such as dropping the genetics requirement, or dropping the virulence and pathogenicity terms can lead to substantial changes in the output list. Although there may well be other approaches that seem plausible and that lead to different lists, the constraint of passing the same statistical tests that we have imposed on our data severely limits the options for new approaches.

A quest for alternatives, however, is probably not worthwhile unless and until the list produced up to this point is assessed by in-depth studies of the literature to which it serves as a guide. The question is not whether we have extended the Geissler list by 10 or 20 new viruses, but rather whether we have extended it at all. In short, we have proposed multiple points of departure for literature studies; even if only a few of them lead to new information that can qualify as a "discovery," the effort will have been worthwhile. In that sense, it does not so much matter whether the list is robust as whether it is useful.

Previous published examples of analyzing complementary literatures have been based on title words and phrases only, not subject headings. It is of interest, therefore, to compare title processing with subject heading as used in the present project. Such a comparison is presented in Appendix III.

12. Significance of Study: A Literature-Based Approach to Scientific Discovery

Whatever merits VL2 has as a virus list, it is first and foremost a guide to small sets of complementary virus literatures in Medline that have two notable features, 1—they are defined by MeSH search terms derived from authoritative criteria for suitability as *bw* agents, and 2—the outcome of the Medline searches, and subsequent processing of downloaded records using Arrowsmith techniques, together indicate a high level of success, as measured by a high degree of statistical significance, in finding viruses of known *bw* interest while finding at the same time new viruses in the same context. Literature structures associated with the new virus terms share important attributes with the Geissler list virus literatures because they were generated by the same search process. The MeSH search terms, their logical combinations, and the Arrowsmith processing may have captured, in scientific, medical, and biological terms, at least some of the meaning stated as military requirements for biological agents that should be considered as potential threats.

The core literatures defined by the initial four Medline searches consist of just a few thousand articles each, very small but important and manageable subsets of the half-million articles in Medline on viruses and virus diseases. A new search that forms the intersection of each virus term in VL2 with each of the Medline-generated sets (*asb*, *air*, *ive*, and *vrlg*) can generate additional subsets that are highly organized by the Medline search terms that created them, and which therefore, merit still higher priority for a focussed study directed toward identifying research that might support the development of *bw* agents. In effect, we have attempted to classify viruses as potential *bw* agents on the basis of their associated literature structures.

13. Next Steps

1. The next and most important step is to assess the value of VL2 by studying in depth the biomedical literature associated with each of its virus terms, bringing to bear appropriate areas of expertise, particularly virology, within the biological sciences. This assessment should include, and perhaps initially should be organized around, the 13 criteria for suitability summarized by Geissler (1986) and quoted in Section 1 of this article. The analysis should be extended beyond that list, especially in the areas of genetics and genetic techniques that might be applied to viruses to alter virulence, pathogenicity, antigenic structure, infectivity, stability, or other characteristics relevant to weaponization (Geissler, 1986). To determine whether G and non-G VL2 viruses are similar in that the state of biomedical research relevant to each may be supportive of weaponization requires evaluation of the core literature that we have identified.
2. Based on step 1, the four Medline search strategies on which the present VL1 and VL2 are based should be reassessed and, if indicated, modified appropriately to produce a new VL1 and VL2. An important aim of any modifications would be to improve the precision of the searches to reduce the labor required to assess the literature that it represents.
3. Extend the study to databases other than Medline, including especially BIOSIS, Scisearch, and Embase. Although all of these have a substantial overlap with Medline, the combination of new modes of access and different scope and coverage (e.g., European coverage in Embase) will enhance the Medline results. The large agricultural databases may also contribute significantly to the scope of the present study, and should be evaluated for that purpose.

Acknowledgment

This work was supported by the Defense Intelligence Agency (DIA) (Contract MDA908-99-M-6679 to The University of Chicago, Don R. Swanson, PI) and by the Office of Naval Research (ONR) with a grant to Neil R. Smal-

heiser, The University of Illinois/Chicago. We are grateful for many valuable suggestions during a series of meetings with a team of text-mining researchers and consultants headed by Dr. Ronald Kostoff of the ONR. This team included, among others, Dr. John Bodnar (DIA), Mr. Chris Centner (DIA), Dr. Charles Clark (DIA), Dr. Michael Lewis (ITT), Dr. Michael Ottlinger (CDC), and Mr. Ray Toothman (RSIS). We are grateful also to an anonymous referee for many valuable suggestions that led to clarifications.

Appendix I: MeSH for Geissler and CDC Lists

To use the Geissler list to test the Medline searches and Arrowsmith procedures, it is necessary first to identify the Medical Subject Headings for each virus on the Geissler list. It is noteworthy that all of the Geissler virus names except three are current species names in MeSH. There are no MeSH terms directly corresponding to the names of the three exceptions, but they are searchable under other names. The appropriate search term was substituted for the Geissler name in the following three cases: first, the Russian spring–summer encephalitis virus is footnoted by Geissler as a subtype of tick-borne encephalitis virus, a note that is consistent with (ICTV, 1996, p. 419). The printed MeSH-alphabetic manual gives as a search instruction "Encephalitis Virus, Tick-borne," which itself is on the CDC list incidentally. Second, the Machupo or Bolivian hemorrhagic fever virus, and the Junin Virus, are members of the Tacaribe Complex (ICTV, 1996, p. 323). The Machupo Virus is indexed with the latter term only. The MeSH search instruction for Junin is to use Tacaribe prior to 1992. For the influenza virus, the MeSH term has changed and is now either "influenza A virus," or "influenza A virus, human." Either term was counted as an acceptable synonym for the same entry.

Certain viruses could be searched only under the genus name rather than the species. Hepatitis A Virus prior to 1992 was indexed with the genus Hepatovirus, and must be searched under that term. Accordingly, the latter was added as a synonymous search term for Hepatitis A virus. Prior to 1994, it was necessary to search Hantaan Virus under Hantavirus; Hantaan virus is the only species given in (ICTV, 1996) corresponding to the Hantavirus genus. Accordingly, the two were treated as synonyms.

As described in the main body of the text, the automatic process for selecting virus terms required a list of all viruses in the defined population. In compiling that list, we omitted 72 terms at the taxonomic level of genus and another 30 terms for virus species for which vertebrates are not the host. The categories we omitted were not part of the Geissler list, and so including them as part of the test criteria would not be justified. Nonetheless, it is of interest to know how these terms would have fared in our search procedures. Accordingly, we separately ran the Arrowsmith procedure for the virus terms that had been eliminated. Those searches that yielded fewer than three records were evaluated. Hepatovirus and Hantavirus were added to VL1, as discussed above. Aphthovirus, Rhinovirus, and Rotavirus were also designated search terms with no subordinate MeSH terms available, and so also were added to VL1. In the evaluation process, we excluded polioviruses, enterovirus, and alphavirus on the basis of examining the resulting abstracts. The airborne category for polioviruses was a spurious search outcome, and the generic alphavirus, arboviruses, enterovirus, flavivirus, and paramyxovirus terms added no information not already represented in the records for the virus species within these genera. Potential additions to rank 4 (seven virus terms) were not evaluated or further considered.

We mention finally the four viruses completely missed in our search process. Three CDC viruses, Sabia, Flexal, and Guanarito are not on the Geissler list and are not MeSH headings (but are indexed under the Tacaribe complex). One more entry, Equine Morbillivirus, is also not on the Geissler list, is not a MeSH heading, and has not been classified by the ICTV, according to their on-line database (see ICTV, 1996, Website), as of 9/3/2000. It recently emerged in Australia, and has been renamed the Hendra Virus. One team researching its genome has suggested that it is not a morbillivirus, and that it be classified as a new genus in the paramyxovirinae subfamily.

Appendix II: Auxiliary Searches

Human Pathogenicity

Once a relatively short list of viruses has been produced (the new terms in the high-rank category), another method of introducing new searches becomes practical without repeating the entire process—namely by searching each of the resulting virus terms to find the intersection with new categories not part of the original search. For example, we briefly explored the idea that human pathogenicity should be included as an important determinant of interest for potential biological weapons.

For each of the virus terms in VL1, a Medline search for pathogenicity (using the subheading py) in records with the check-tag "human" was conducted. There were 14 terms for which only 0, 1, or 2 records resulted, none of these being Geissler terms. All of the Geissler terms had four or more records. The 14 terms are marked in the VL1 raw data list (Table 1) in the third column, where the number of records (0, 1, 2) is shown for these terms. Examining the records for the terms with one or two records indicated that the association between the index terms py and human was spurious, and so these were marked with an "x" to denote that observation. Because of our interest in detecting viruses that might have been genetically altered, each of the 14 terms was also searched for any cooccurrences with the subject heading "genetic vectors"; all except two (Infectious peritonitis virus feline and Kemerovo virus) had one or more such cooccurrences. For this reason, all of these terms were retained; the "x" markings are for information purposes

only, thus leaving to the user the decision of whether to keep, discard, or further investigate these terms.

Genetic Techniques

The first Medline search, *vrlg*, involved the subheading "genetics (ge)," but did not explicitly include "genetic techniques" (a section of MeSH separate from "genetics"), a subdivision of which is genetic engineering. These latter categories are definitely of interest, so a similar procedure to that above for pathogenicity was followed for "genetic techniques." Many records were found for each virus entry, both for the G-terms and the non-G terms. It appeared, therefore, that this subject category was well represented in the original *vrlg* search, and afforded no basis for further distinguishing G from non-G terms. (This outcome was not unexpected; the genetics and genetic techniques categories overlap heavily within the Medline database.)

Appendix III: Comparison with Arrowsmith Title-Word Searching

Arrowsmith is freely available for public use at http://kiwi.uchicago.edu or at http://d-swanson.uchicago.edu. This Web version is presently (3/2001) based on title words and phrases only, and ignores subject headings. It is of interest, therefore, to compare title processing with subject headings as used in the virus project.

The three sets formed with subject heading searches (*asb*, *air*, and *vrlg*) were submitted in two pairs (*asb* with *vrlg* and *air* with *vrlg*) to Arrowsmith for title processing. Normally, Arrowsmith input should be based on title-word searching, but in this case we want to compare identical sets processed in two different ways—hence, exactly the same sets of records (for the airborne cases only) were submitted for title processing as were used for the subject heading procedure; that is, the same Medline search strategies were used in the two cases.

The two search pairs, *asb-vrlg* and *air-vrlg* produced two B-lists, which were then edited and compared with ranks 1, 2, and 3 of VL2, which were based on subject headings. Rank 4 was ignored because all of the terms are from the set *ive*, which was not a participant in the title run. The results are in Table 3. However, the number of non-G terms is only an estimate because of the difficulty, when editing title-word B-lists, of distinguishing a virus from a viral disease. The mention of a disease does not imply that the article is about the virus, yet an article about both the virus and its disease may mention only the disease in its title. (In the subject heading procedure, only virus subject headings were used, not the disease headings.)

In any event, it is clear that the title search is reasonably good at reproducing most of the terms in VL2, but with many additional terms as well, possibly so many as to deprive the complete list of any statistical significance in terms of the density of G-terms. This conclusion could not have been materially altered by including the better-performing *ive* case, because the additional non-G terms would still be present and the G-terms were already well represented. (Statistical significance for the title B-list is almost meaningless in any case because of the heavy reliance on manual editing. Any list can become statistically significant if enough non-G terms are manually deleted.)

Title processing would not have been as effective as subject headings in producing a statistically based ranked list such as VL2. However, now that VL2 has been produced, title processing should be valuable for further exploration of the related literature.

References

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

Anderson, W.C., III, & King, J.M. (1983). Vaccine and antitoxin availability for defense against biological warfare threat agents. U.S. Army Health Care Studies Division Report No. 83-002. Fort Sam Houston, TX: US Army Health Services Command.

Bates, M.J. (1999). The invisible substrate of information science. Journal of the American Society for Information Science, 50(12), 1043–1050.

Carus, W.S. (1991). "The Poor Man's Atomic Bomb?" Biological Weapons in the Middle East. Policy Papers Number 23. Washington, DC: The Washington Institute for Near East Policy.

Everitt, B.S. (1977). The analysis of contingency tables. New York: Wiley.

Geissler, E. (Ed.). (1986). Biological and toxin weapons today. Oxford, UK: SIPRI.

Harter, S.P. (1984). Scientific inquiry: A model for online searching. Journal of the American Society for Information Science, 35(2), 110–117.

ICTV. (1996). International Committee on the Taxonomy of Viruses. Sixth Report. Virus Taxonomy. http://www.ncbi.nlm.nih.gov/ICTVdb.

Minitab Data Analysis Software. (1998). Version 12. State College, PA: Minitab, Inc.

National Library of Medicine; (NLM). (1999). Medical Subject Headings (MeSH)—Annotated Alphabetic List; Permuted Medical Subject Headings; Medical Subject Headings—Tree Structures; http://www.nlm.nih.gov/mesh/filelist.html.

Roberts, B. (Ed.). (1993). Biological weapons: Weapons of the future? (vol. XV, number 1). Washington, DC: The Center for Strategic and International Studies.

Smalheiser, N.R., & Swanson, D.R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. Neuroscience Research Communications, 15, 1–9.

Smalheiser, N.R., & Swanson, D.R. (1996a). Indomethacin and Alzheimer's Disease. Neurology, 46, 583.

Smalheiser, N.R., & Swanson, D.R. (1996b). Linking estrogen to Alzheimer's Disease: An informatics approach. Neurology, 47, 809–810.

Smalheiser, N.R., & Swanson, D.R. (1998a). Calcium-independent phospholipase A2 and schizophrenia. Archives of General Psychiatry, 55, 752–753.

Smalheiser, N.R., & Swanson, D.R. (1998b). Using Arrowsmith: A computer-assisted approach to forming and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine, 57, 149–153.

Swanson, D.R. (1986a). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7–18.

Swanson, D.R. (1986b). Undiscovered public knowledge. Library Quarterly, 56, 103–118.

Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science, 38, 228–233.

Swanson, D.R. (1988). Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine, 31(4), 526–557.

Swanson, D.R. (1989a). A second example of mutually-isolated medical literatures related by implicit unnoticed connections. Journal of the American Society for Information Science, 40, 432–435.
Swanson, D.R. (1989b). Online search for logically-related noninteractive medical literatures: A systematic trial and error strategy. Journal of the American Society for Information Science, 40, 356–358.
Swanson, D.R. (1990a). Somatomedin C and arginine; Implicit connections between mutually-isolated literatures. Perspectives in Biology and Medicine, 33(2), 157–186.
Swanson, D.R. (1990b). The absence of co-citation as a clue to undiscovered causal connections. In C.L. Borgman (Ed.), Scholarly communication and bibliometrics (Ch. 7, pp. 129–137), Newbury Park, CA: Sage.
Swanson, D.R. (1990c). Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association, 78, 29–37.
Swanson, D.R. (1991). Complementary structures in disjoint science literatures SIGIR91. In A. Bookstein et al. (Eds.), Proceedings of the fourteenth annual international ACM/SIGIR conference on research and development in information retrieval, Chicago, Oct 13–16, 1991 (pp. 280–289). New York: ACM.
Swanson, D.R. (1993). Intervening in the life cycles of scientific knowledge. Library Trends, 41(4):606–631.
Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence, 91, 183–203.
Swanson, D.R., & Smalheiser, N.R. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. Library Trends, 48(1), 48–59.
U.S. Senate. (1998). Hearing. Committee on the judiciary, biological weapons; The threat posed by terrorists (pp. 22–23).
U.S. Code of Federal Regulations. (1999). 42 Parts 1–399 Public Health (Ch. 1 10-1-99 ed.; Part 72.7 Appendix A. pp. 528–529).
USA. (1964). Military biology and biological agents. Departments of the Army and the Air Force (March), Technical Manual TM 3-216/AFM 355-6.