ADVERSE EVENT & DISEASE REPORTING, SURVEILLANCE, AND ANALYSIS

A DIMACS Working Group

DIMACS
The Center for Discrete Mathematics and Computer Science

Summary of 1st Meeting (Oct. 16-18, 2002)

The meeting program was designed to have presentations covering a range of Adverse Event / Disease surveillance issues including its purpose, the data sources currently available and analytic approaches. The schedule (see http://dimacs.rutgers.edu/Workshops/AdverseEvent/program.html) was designed to (1) first present individual overviews for the purpose of establishing common communication vehicles among participants, (2) to then facilitate interaction among participants through working subgroups as soon as the overview presentations were accomplished (3) to next gather feedback from participants (working subgroups) and establish priority themes that would be approachable within the DIMACS forum and next, (4) continue with individual presentations on various data sources and analytics, while (5) interweaving discussions of specific follow-up activities for the working group and (6) finally establish a means to mobilize the specific follow-up activities that could best be accomplished by taking advantage of the working group momentum.

The participants comprised a range of interests and disciplines related to surveillance including: medical epidemiologists, local health department officials, academic leaders, federal officials from regulatory (drugs and biologicals) and public health policy agencies, statisticians, mathematicians, transactional data system developers, analytic data manager/developers, and computer scientists.
(see http://dimacs.rutgers.edu/Workshops/AdverseEvent/participants.html)
Differences in interests were evident almost immediately. There was a perceived dichotomy with each category forming two conceptual branches as described in Figure 1.

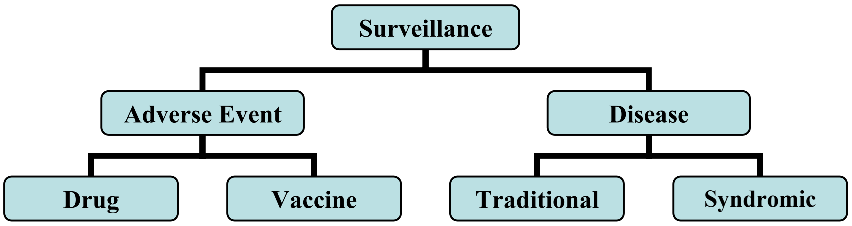

Figure 1

Observing and acknowledging this in the meeting helped identify and apply attention to aspects of commonality across the differing branches of Figure 1. Examples of the common concerns for all branches in Figure 1 included information systems integration, data standards, administrative and other spontaneous data usage, analytic methods and their interpretations, privacy/confidentiality/security etc. The commonalities of concern began to crystallize within the group process explicitly after the breakout sessions of the first day. The organizers began trying to maintain focus on the regions of overlap shown in Figure 2, for the purpose of exploiting the working group's potential.

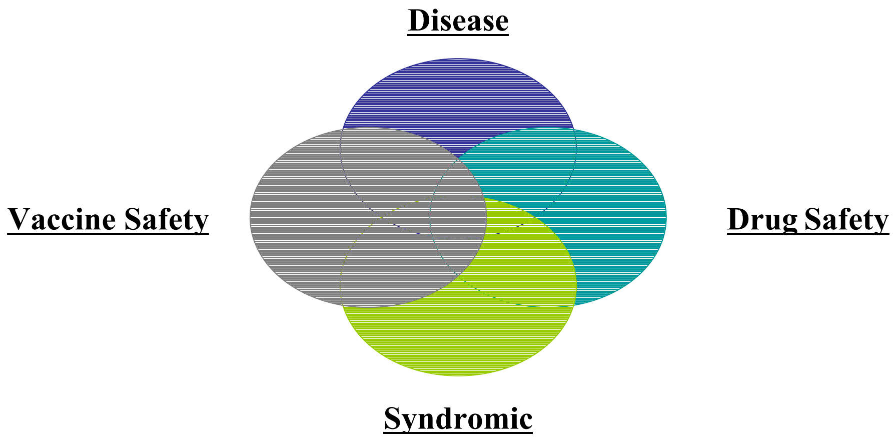

Figure 2

There were 5 (breakout) working subgroups each with a designated facilitator and recorder. The working subgroup format was designed to familiarize participants with one another by addressing vital components necessary for successful progress in this meeting. The topics addressed were:

1. *What improvements could be made in the national (international?) capability of developing new and better analytic tools (software) and methods for analysis of Adverse Event surveillance, etc.?*

2. *Concerning disease and adverse event reporting and surveillance, what are the major issues that need to be addressed? Groups should be cognizant of peace-time needs as well as deliberate threats.*

3. *Develop a list of research priorities and recommendations in the area of detection and analysis of Adverse Events, etc. Related issues include: the role of government, the role of higher education, software development and privacy.*

4. *How do we stitch multilevel, multiple stake-holder surveillance into a coherent whole?*

5. *How can the working group format used at this conference serve a useful role? What format should future working group activities adopt?*

After the breakout sessions, the entire group reconvened to present and discuss the conclusions of each working subgroup.   See appendix A for the notes from the discussions.

Although the actual program did not progress *exclusively* from general overview to more specific presentations, that was the general direction. Guided discussions focusing on what and how to plan for follow up activities within DIMACS and the working group were interspersed between presentations. A number of themes developed. These included:

- Leadership
- Standards
- Case/AE definitions
- Information exchange
- Timeliness for response potential
- Data access
- Data quality factors
- Signal detection thresholds
- Testing Methodologies
- System Evaluation

Discussions led to consideration of what may and may not most appropriately be addressed in the DIMACS working group forum.   Given the mathematical and computer science orientation of the collaboration, it seemed most logical to try and concentrate working group activities on components of the surveillance picture that relate most closely to analytic applications, their evaluation, data preparation and interpretation of results *from* the analytic signal detection applications. (See figure 3) There is a logical conceptual flow through the components of the diagram (i.e. the arrows) and although none of the components are functionally independent, there are areas of inextricable connectivity to analytic applications. These are shown by broken borders in the diagram.

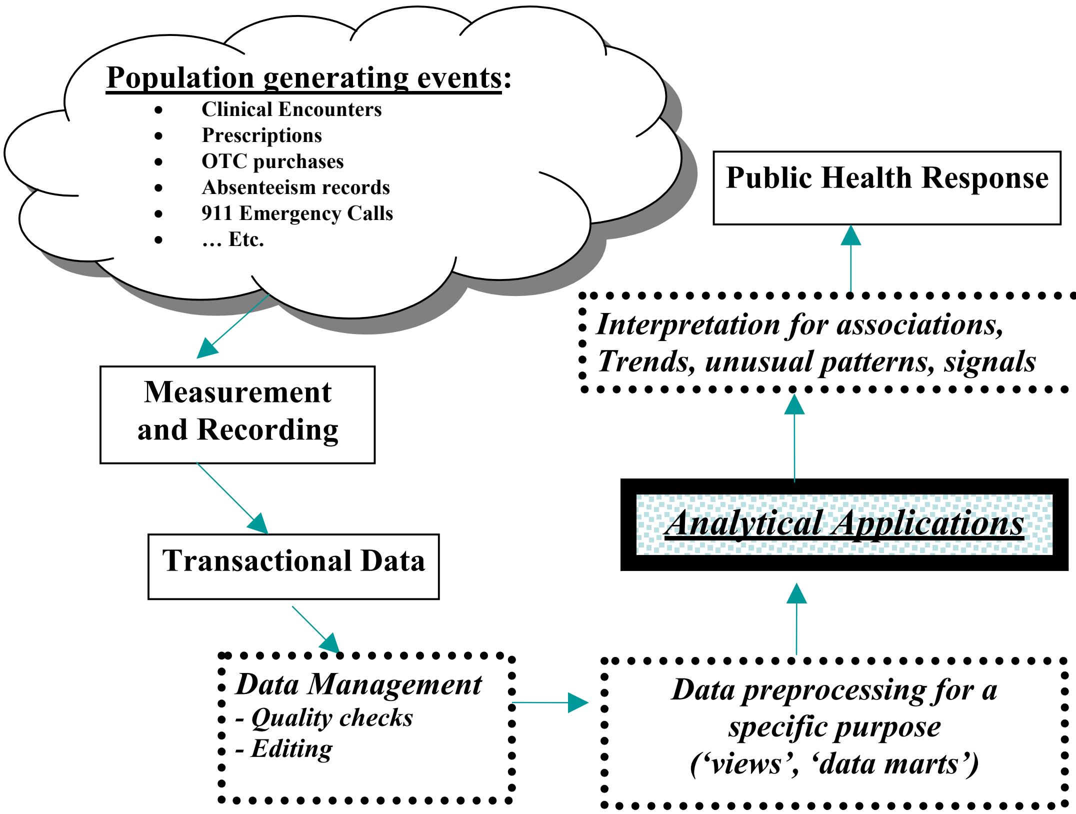

Figure 3

Since much of the data used for public health surveillance is not collected specifically for that purpose and/or is spontaneously generated, (1) the data require substantial pre-processing, (2) a sample-to-population mapping is not probabilistically defined and therefore, (3) the existing analytic signal detection methodologies that could be applied to surveillance data are empirical in nature and do not lend themselves to conclusions bearing well-defined confidence intervals or p-values. This 3rd aspect has not received a substantial degree of productive attention. Little has been established in public health academic curricula to incorporate use of realistic data to develop updated methodologies to meet new surveillance requirements.

The primary users of surveillance systems in government have the practical needs and applied experience but do not have the resources, nor are they structured to perform highly technical research and development as are academic institutions and the private sector. For these reasons, many signal detection methodologies that are currently in academic use have not been well characterized for their application value to existing data. There is a critical need to more fully characterize the quality of conclusions from existing data that lead to interventions, programmatic and policy decisions. It is impossible to evaluate surveillance systems in general without a means to rigorously evaluate all of the components. The development, refinement and practical characterization of analytic methodologies thus represent a weak link in surveillance that could be strengthened by activities arising out of this working group. This is a guiding principle for the future activity recommendations proposed for this working group.

The **main goals** of the follow-up recommendations are to:

A. Foster the development of analytic signal detection methodologies and algorithms that perform successfully in realistic data structures and surveillance systems.

B. Establish and maintain a forum for evaluating the operating characteristics of signal detection methodologies through simulated and if possible real data.

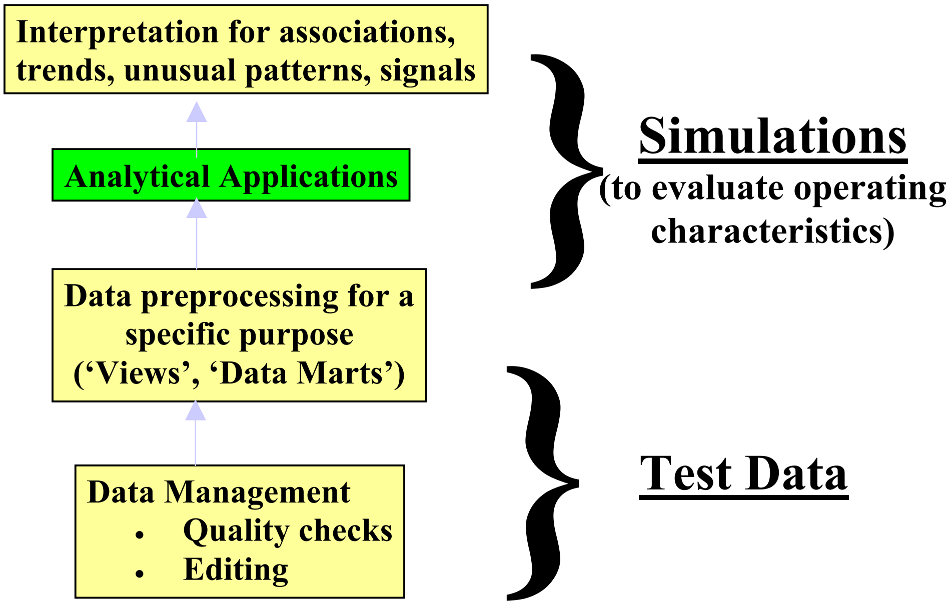
Interpretation for associations, trends, unusual patterns, signals
Analytical Applications
Data preprocessing for a specific purpose ('Views', 'Data Marts')
Data Management
• Quality checks
• Editing
Simulations
(to evaluate operating characteristics)
Test Data

Figure 4

Some **activities, tasks, outcomes and/or requirements** for fulfilling these goals are:

- Development of realistic, modern test data sources for refining signal detection procedures.
- Development of important "signals" to introduce into the data sources.
- Making the data available for refining algorithms.
- Establishing a competitive forum (timing, rules, judging, etc.) to test developed methodologies.
- Develop related simulations to assess operating characteristics.
- Develop flexible algorithms for mimicking data base characteristics.
- Produce modular, interoperable signal detection algorithms.

- Obtain or estimate computing efficiencies for Monte Carlo simulations of signal detection events in large complex data.
- Use multidimensional graphical displays to communicate/interpret results, and evaluate algorithms.
- Apply multivariate statistical techniques for evaluating signal detection profiles across multiple data sources.

Recommendations resulting from the first meeting of the *Working Group on Adverse Event/Disease Reporting, Surveillance and Analysis* are to continue pursuit of this subject in a more focused manner by:

A. Establishing a working subgroup to initiate and conduct a signal detection competition similar to the Text Retrieval Conferences (TREC; see http://trec.nist.gov), using real and/or realistic surveillance data. The competition would build on the series of DIMACS "algorithm implementation challenges" (see http://dimacs.rutgers.edu/Challenges). A side effect of this activity would be to create a data repository for research and experimentation with surveillance methods. A second side effect would be to catalog the available tools and highlight specific needs for tool development.

B. Establishing and maintaining a working subgroup to evaluate by simulation the operating characteristics of analytic applications for detecting meaningfully defined and specified signals of importance.

C. Forming a Working Group focused on privacy, confidentiality, and security issues associated with healthcare data. This group will explore the uses of relevant modern cryptographic methods, a particular strength at DIMACS. Privacy concerns are a major stumbling block to public health surveillance, in particular bioterorrism surveillance and epidemiological research. How to detect patterns in large data sets or share information between data sets while maintaining privacy and confidentiality of the data is a serious challenge. The problem is of interest to government agencies at all levels of government, industrial and academic researchers, as well as to a growing commercial sector that collects, maintains, and markets such data sets. Not many computer scientists knowledgeable about methods of cryptography/security/privacy/ cryptography have become involved in this area and the area is ripe for new partnerships between persons in the public health/ epidemiology community, the health data industry, and the computer science community. This working group will develop such partnerships.

D. Developing a summer school or week-long tutorial on statistical methods for surveillance

E. Developing a monograph or special issue of a journal to promulgate the Working Group's activities and stimulate research activities.

F. Preparing a detailed report of research challenges in the Adverse Event and disease surveillance area.

G. Exploring internships/partnerships between researchers and State/Local Health Departments

With regard to recommendations A and B, we are planning to convene a meeting sometime in January 2003 to formulate an action plan.

Don Hoover, Rutgers University
David Madigan, Rutgers University
Henry Rolka, CDC

Appendix A

Reports of Working Group Facilitators

Panelist #1 – Chan Russell of Lincoln Technologies

What improvements could be made in the national (international?) capability of developing new and better analytic tools (software) and methods for analysis of AE, etc.?

This group was comprised of Jana Asher, Robert Ball, Martin Kulldorff, Chan Russell, John Stultz, and Juhaeri. The group was diverse, with folks having expertise in adverse events, pharmaceutical issues, and in disease surveillance. First Chan invited Martin Kulldorff up to list three areas for improvement of analytic tools. These areas were:

1. Text identification tools, in order to match a particular event described in different ways in different datasets. An example of where this would be difficult from the NYC data was the E.R. description of a patient "not being himself."

2. Elimination of duplicates within a dataset in an automatic fashion. Even if a portion of this is done by hand, as much as possible should be done automatically.

3. Signal detection. Martin pointed out that there are three basic types of data: temporal, geographical, and database, and that each of these types could be used for either retrospective or prospective analysis. Martin clarified that database data contains information on age, gender, etc… basically, covariate information, while temporal and spatial data just contain reports of adverse events by time period or by place. While there are already a lot of methods for analyzing temporal and spatial data, the prospective analysis of spatial/temporal data is in its infancy.

Martin pointed out that it is important to make adjustments for known relationships in the data, and to adjust odds for things we know about (e.g., side effects of a drug that are known).

Finally, there was a discussion of how to provide tools/software to users and collectors of data by the group. Jana suggested the concept of the "grey box." A black box is opaque to the user, but a grey box is moderately simple to use, but required a small amount of training for the user and the user has some basic knowledge of how the box operates. This seems like a reasonable compromise between the users' desire to have a simple, "push a button" analysis package, and the experts' concern that the tools they develop are used properly.

At this point, Chan spoke again and added a few small points to Martin's discussion.

1. He mentioned the need data cleaning, and gave the example of the drugs reported in AERS as having several versions of their names used.

2. There are existing tools the group discussed that would be useful, such as name standardization and record linkage software used at the U.S. Census Bureau, and a SAS text data miner mentioned by John Stultz.

3. Chan pointed out that he really liked grey box idea, suggesting the dissemination of a software package with tutorials, training and support. This would be quite different than a neural net black box.

4. Chan pointed out that there is a lot of problem in signal detection in low quality data, and that improving the quality of data has to be a priority. He discussed ways better (more standardized) data could be collected, perhaps by developing tools and distributing such tools to states and local data collectors. Also, the Internet could be used to improve data collection at the source of the data.

5. He also talked about the need to do the analysis but then sort through the data and be careful about presentation of the results of an analysis to non-technical people (i.e., go back to graphs). He suggested that this connects into the grey box idea; suggestions for data presentation would be involved in created the grey box.

6. A final topic discussed by the group was capture-recapture. While it can be useful, there are challenges in getting enough information out of databases to know you are talking about the same people in each list.

Panelist #2 – David Walker of CDC

Concerning disease and adverse event reporting and surveillance, what are the major issues that need to be addressed? Groups should be cognizant of peace-time needs as well as deliberate threats.

This group consisted of David Banks, Teresa Hamby, Jane Harman, Meade Morgan, and David Walker. The group has members from federal, state, and international levels. They focused on what can be done to get this working group contributing at a state/local level. The notes presented consisted of four major topics.

1. The role that DIMACS should play. The group suggested several workshops:
   a. A workshop to develop functional specifications for standardized bioterrorism detection.
   b. A workshop to explore exposure (Bayesian) estimation versus standard *Proportional Risk Ratios* for risk assessment.
   c. A workshop to identify software tools to mix and match, addressing such needs as scan statistics, multiple systems analysis, record linkage, data streaming, GIS, and data quality.

2. Identifying analytical needs and developing capabilities. The group identified four separate topics with different needs:
   a. Bioterrorism needs speed.
   b. Drug events need rich data.
   c. Safety monitoring needs exposure estimates, data quality checking, etc.

d. Surveillance systems (environmental, chronic illness, criminality, congestion) need analytic methods that are application specific.

3. Mathematical techniques to facilitate secure sharing of information across institutional boundaries and jurisdictional barriers.

4. Facilitating NEDSS, AERS, VAERS.

Discussion of Panelist #2's Presentation

The question arose as to who determines if NEDSS data meets quality requirements. David Walker answered that for an outside entity to do this would be a good idea, since internal review can be subjective. Dan Sosin pointed out that an independent contractor has evaluated NEDSS. How analytic tools are related to NEDSS would be a subtopic of a review. Dan also pointed out that NEDSS has undergone internal critical review as well. It was suggested that something independent/more careful has to be done. David Walker suggested more syndromic surveillance.

Chan Russell brought up the research topic of cryptographic techniques for identifiers as a balance between privacy considerations and the need to match across diverse sources. Dan Sosin asked who may already do research on this. He stated that someone is already exploring such tools, and that DIMACS should encourage these existing efforts. Someone else mentioned that confidentiality studies are being done, where individual records in datasets are perturbed, estimates are created from the modified datasets, and then these estimates are compared to true estimates from the original datasets. David Madigan mentioned someone will be talking about confidentiality tomorrow. Fred Roberts mentioned that this topic will fuel a breakout session, and asked if the working group recommended that a subgroup be devoted to this topic. David Banks suggested the topic for the subgroup be expanded to include personal key infrastructure in general.

Panelist #3 – Dan Sosin of CDC

Develop a list of research priorities and recommendations in the area of detection and analysis of AE, etc. Related issues include: the role of government, the role of higher education, software development and privacy

This group included Coleen Boyle, Owen Devine, Joe Fred Gonzales, Dan Sosin, and Ana Szarfman. The group determined four categories of research priority areas, which follow.

1. Collection of data
   The group pointed out that health data lacks standardization, and suggested case definitions require validation and automation. IT tools are required to increase timeliness of data collection.

2. Management of data
   Record linkage techniques need to improve, in terms of person-record matching and geo-referenced data. The group discussed cleanliness of data versus timeliness of

data as a trade-off and an issue. They suggest IT tools to create analysis files and increase efficiency of large database performance.

3. Analysis of data
   Modeling and simulation studies and research are required. Topics for research include:
   a. Aberration detection and standards for pattern recognition to prompt investigation.
   b. Exploration of relationships between novel data (i.e., GIS, Veterinary) and health outcomes.
   c. Development of automated analytic systems (e.g., AI/neural networks), perhaps that use historical data.

4. Interpretation of data – the group suggests developing metrics of evaluation to compare and contrast approaches and tools (on a side-by-side basis), in order to prioritize and develop tools for assessment and analysis (using usefulness, sensitivity, timeliness all as rating factors).

The group also brought up a few other issues. First, there are jurisdictional issues, in the sense that much data is collected on the local level and therefore there is variability in the data when merged into state or federal level systems. Harmonizing from the local to the state to the federal level is therefore an issue, and barriers to standardization must be determined so that solutions can be developed. Also, sentinel surveillance, in terms of generalizability and the role of sampling and/or targeting populations should be explored. Finally, the dual use of information systems should be addressed; systems should be useful to both the clinicians that enter the data and the public health.

Discussion of Panelist #3's Presentation

Ana Szarfman pointed out the importance of removing bottlenecks. She also commented that putting together disease surveillance and adverse effects as the subject for one working group is awkward. Fred Roberts pointed out that one recommendation his group made is to expand on the list of research challenges and produce a report of these challenges. He then asked if it is a reasonable objective to produce such a list of research challenges, and asked what the working group thought would be useful. Robert Ball suggested processing the notes from these discussions and compiling them, and then distributing them for group comment. G. David Williamson suggested labeling the importance of each research issue, and determining what challenges are long-term, and what can be resolved short-term. Fred Roberts pointed out that the DIMACS bioterrorism group created a similar report, and that they found prioritizing what to do difficult. He suggested simply starting with a list of topics and getting volunteers to put 1-2 sentences to each topic. Another working group member, a data user, stated that prioritizing research challenges would be very helpful to her. Dan Sosin suggested having the working group create a list of research needs. Another working group member suggested defining priorities was essential. Henry Rolka suggested the working group work on this tomorrow at 3:40pm. Finally, G. David Williamson stated that he thought 3 levels of priority on research challenges are enough to do the job.

Panelist #4 – Donald Hoover of Rutgers University

How do we stitch multilevel, multiple stake-holder surveillance into a coherent whole?

This group contained Miles Braun, Donald Hoover, Mayron Katzoff, Wendy Wattigney, and Farzad Mostashari. The group decided it was more appropriate to think about "what to" not "how to", and reported having a very animated discussion about the topic. Donald presented a list of points/questions compiled by the group.

1. How does Adverse Event Surveillance fit into the larger role of taking remedial action? There are legal issues related to monitoring adverse events… what is the point of having a complex algorithm to do something that you can't act upon?

2. What consideration is needed for integrating local systems into a national system?

3. What do you do about false alarms? The cost of false alarms may be prohibitive, and there may be legal issues. There therefore needs to be quality control/piloting of systems so that a "bad method" isn't put into practice.

4. There is a question of the relationship between research methods and public health practice. The needs of public health practice should drive the research, but research is often divorced from public practice and presented as public practice in order to get research money. An analogy between republican/democratic primaries (where the democrats make themselves more liberal and the republicans more conservative during the primaries, and then move toward the middle later) and the research/practice tension was made – those in jobs either do research or do practice, hard to come together in the middle.

5. Data standardization is needed. $918 million is given by the federal government to states for the states to develop systems to monitor bioterrorism on their own. They aren't working together, so there is no data standardization. This will affect the goals of this DIMACS working group.

6. Research methods need to be developed. There need to be quality control of methods; for example, new methods can be tested by using them to attempt to find a signal in historic data. An example of the problems that can occur with methodological development was given; there was a project where 12 different modeling procedures were applied to data on AIDS cases in London between 1989 and 1993. All 12 modeling procedures produce different predictions of future AIDS cases. This is an issue of quality control.

7. Funding of research as an issue came up; such funding would probably come from the government. Other related issues include how to find people to do research, academic versus corporate research, and that industry may not be the best place for research right now.

8. Bringing new researchers into the field is another issue. How do we find these people (e.g., Statistics Ph.D.'s and Computer Scientists) and generate interest? A suggestion for an academic bioterrorism center (cartel) was made, to promote analytic research in disease surveillance.

9. Finally, existing analytic problems were discussed. There is a need to develop computationally tractable methods, with good aberration detection within health data, multiple stream data, real-time data, and spatial data. The methods need to work better than the eyeball! Another issue is a lack of definition of syndrome groups. There is a field called medical ontology (classification of diseases) that should be drawn upon. The issue of multiple comparisons needs to be addressed as well.

After compiling these points, the group discussed where DIMACS could play a role. They felt DIMACS could focus on analytics, training, and software development and testing. There may be the opportunity for collaboration with health departments if such department is willing, and matchmaking with health departments and government agencies would be useful. DIMACS could promulgate standards for analysis, and could promote applied research centers and consortiums.

Discussion of Panelist #4's Presentation

The question was raised of how to stitch everything mentioned into a coherent whole… and whether Donald touched on that charge. A participant mentioned that his group talked about some of the same things Donald discussed.

The question was raised by as to whether someone could have used an analysis technique and have found AIDS faster than the eyeball method? Ana Szarfman gave an example of the importance of having ancillary information; she researched a question about Multiple Sclerosis on Google, and found a connection between M.S. and Cylert that she couldn't have found without ancillary information. The contrast between the eyeball method and algorithmic tools was discussed, and that a good compromise is to use the analytic tools, create graphs, and then eyeball, and that it doesn't need to be an either/or situation. Someone gave the example of chess; computers have not been good chess players until recently. Finally it would be good to summarize success stories on the usefulness of analytic techniques.

Panelist #5 – David Madigan of Rutgers University

How can the working group serve a useful role? What format should future activities adopt?

This group included David Madigan, Rick Picard, Fred Roberts, Henry Rolka, and Leslie Todorov. They first addressed the question of how the DIMACS working group can play a role in future research. They came up with the following purposes for the DIMACS working group:

1. Stimulating methodological research and development
2. Initiating training programs
3. Generating communications among data owners, decision makers, and researchers
4. Facilitating data access for researchers
5. Developing evaluation of methodologies
6. Disseminating developments to data users and developers

The group's concrete suggestions for implementing these purposes were:

1. An algorithm challenge/competition
2. A report on research challenges
3. A summer institute or short course
4. Interaction with the DIMACS data mining working group
5. Creation of working group subgroups on the following topics:
   a. Text mining
   b. Data quality
   c. Data streams
   d. Security/crypto
   e. Evaluation standards for analytic methods
   f. Analytic methods for signal detection across multiple data source/types
6. Creation of a data repository
7. Creation of internships and/or partnerships with health depts.
8. Publication monograph or journal special issue

Discussion of Panelist #5's Presentation

During the presentation of the purposes for the DIMACS working group, there was a request by Dan Sosin for clarification of what this DIMACS working group is and in what context it is working. David Madigan felt that the list of concrete suggestions was the best response to this question, and proceeded with his presentation.

After David presented the concept of an algorithm challenge or competition, a discussion of what this would entail ensued. David described the algorithm challenge as the creation of a realistic high-fidelity large-scale dataset coupled with a challenge to researchers to mine the dataset. Dan Sosin asked if what was given in the challenge was just a dataset or if there were scenarios presented with the data. David Madigan stated that there would be a diverse range of problems presented within different tracts within the competition. G. David Williamson pointed out that this would promote data access as well as the development of analytic methods. Henry Rolka stated that data made available for this would not be a simple flat-file textbook-like example but a multidimensional, longitudinal database of realistic proportion and quality.

Fred Roberts mentioned that DIMACS has done 7-8 of these competitions so far. Basically, a committee would be designated that would create a challenge problem. Researchers that took up the challenge would be given intermediate feedback by the committee while they worked on the challenge. At the end, hopefully better algorithms will have been created, and the creators of the better algorithms would be invited to a conference to present their work and exchange ideas.

Dan Sosin asked what it takes to build a challenge dataset. Chan Russell suggested such a dataset would be created from an artificial database or a database like the VAERS or AERS databases, into which a signal would be injected. Henry Rolka suggested starting with the VAERS database and using it as background noise. Then a signal would be injected; the whole procedure would be complicated. David Madigan asked if this is what we really want to do, given that it is a non-trivial undertaking. He wanted to know what the price tag of such an endeavor would be, and suggested a subgroup would have to look at that. Martin Kulldorf mentioned that in NYC, benchmark datasets were already being developed. They haven't figured out how to disseminate these data yet, but the point is to try to use NYC baseline data and inject different types of outbreaks. He said we could discuss the project with him.

The question arose as to whether this activity would target more than one particular type of system? Fred Roberts pointed out that the fairly focused challenges have been more successful. He suggested that we might do more than one challenge, and pointed out that DIMACS has always done these challenges on a shoestring budget. The biggest resource required is the time of a volunteer committee. Finally, Miles Braun pointed out that generalizability is a big issue for such a challenge, and that we should carefully consider this.