行政院所屬各機關因公出國人員出國報告書

（出國類別：出席國際會議）

# 參加「2003 年第二屆國際症狀監視系統研討會」出國報告書

## （2003 National Syndromic Surveillance Conference）

服務機關：行政院衛生署疾病管制局

出國人 職稱：科長、助理研究員、科員

姓名：張啟明、吳芳姿、林麗雯

出國地區：美國

出國期間：92 年 10 月 19 日-10 月 27 日

報告日期：93 年 1 月 14 日

系統識別號:C09203590

# 公務出國報告提要

頁數: 12　含附件: 否

報告名稱:

赴美參加2003年第二屆國際症狀監視系統研討會

主辦機關:

行政院衛生署疾病管制局

聯絡人／電話:

黃貴玲／23959825x3022

出國人員:

熊昭　行政院衛生署疾病管制局　顧問
張啟明　行政院衛生署疾病管制局　資訊小組　主任
吳芳姿　行政院衛生署疾病管制局　研究檢驗組　助理研究員
林麗雯　行政院衛生署疾病管制局　疾病監測調查組　科員

出國類別: 其他

出國地區: 美國

出國期間: 民國 92 年 10 月 19 日 -民國 92 年 10 月 27 日

報告日期: 民國 93 年 01 月 14 日

分類號/目: J4／公共衛生、檢疫　／

關鍵詞: 症狀監測系統，急診資料監測，即時監控

內容摘要: 參加於New York Academy of Medicine為期2天的Conference，及研討會前為期3天的Workshop課程。了解美國各州目前進行症狀通報系統（含急診室監測及藥房通路監測）發展情形，及學習系統分析使用之應用程式；依美國進行時所遭遇之困難及進展，檢討目前台灣應該進行發展監視系統的方向。

# 摘　要

參加於 New York Academy of Medicine 為期 2 天的 Conference，及研討會前為期 3 天的 Workshop 課程。了解美國各州目前進行症狀通報系統（含急診室監測及藥房通路監測等）發展情形，及學習系統分析使用之應用程式；依美國進行時所遭遇之困難及進展，檢討目前台灣應該進行發展監視系統的方向。

# 目　　次

## 壹、目的

台灣自SARS疫情爆發後，傳染病監視系統之重要性，實不容忽視，目前已有法定傳染病、症候群、定點醫師、學校、院內感染等監視系統，但多數為被動監視，均依賴通報者（如：醫師）的診斷及通報意願。SARS疫情期間，美國疾病管制局衛生人員曾來我國協助疫情，紐約州的衛生人員曾告知，紐約州近三年來正進行一項新的監視系統，從自願參加之公立醫院急診室，收集來診個案之症狀，依症狀別進行長期監視，用以即時偵測疫情。本局為了解國外現階段的做法，故參加New York City Department of Health and Mental Hygiene 於92年10月20至24日舉辦之第二屆全國症狀監視研討會（2003 National Syndromic Surveillance Conference），以提供本國未來建立新的疾病監測系統之參考。

## 貳、過程

此次出國行程自92年10月19日起至10月27日止，含路程所需時間共計9天。除參加於New York Academy of Medicine為期2天的Conference，並參加研討會前為期3天的Workshop課程。Workshop課程內容及研討會議題及相關資料如附件一、附件二。

行程如下：

| 時 間 | 地 點 | 紀 要 |
|---|---|---|
| 10/19 | 台北→美國紐約 | 路程 |
| 10/20 | 美國紐約 | 抵達 |
| 10/21 | 美國紐約 | Workshop |
| 10/22 | 美國紐約 | Workshop |
| 10/23 | 美國紐約 | Workshop |
| 10/24 | 美國紐約 | Conference |
| 10/25 | 美國紐約 | Conference |
| 10/26 | 美國紐約→台北 | 路程 |
| 10/27 | 台北 | 抵達 |

## 參、心得

第二屆症狀監測系統研討會於美國 New York City Department of Health and Mental Hygiene 舉行，400 多位各國公共衛生人員、醫師、專家學者（包含美國聯邦政府軍醫人員）共同討論在常規的監視系統外，收集無診斷（症狀）之資料，並利用電腦系統，建立資訊化的監視系統，偵測異常狀況，以監控疫情發生。

美國目前監測系統包含：疫情監測系統、早期警示系統、健康指標監測系統、症狀監視系統等。

由於國內尚未建立「症狀監視系統」，故這次參加研討會，主要欲瞭解美國及其它國家實施現況，以及過去及目前曾遭遇的問題、系統所使用之資訊模式等。簡述如下：

一、症狀監視主要的工作有四大部分

1. syndrome definition（10% ）
2. information（30% ）
3. statistic model（30% ）
4. evaluation（30% ）

二、雖然美國 Syndromic surveillance system 才剛開始，但政府衛生單位積極進行有系統性的開放發展，因此，在此次會議中可以見到各州衛生單位及學界均投入相當的人力進行系統架購，及臨床資料分析，成效及發展速度皆相當的快速。

三、注重與資訊的結合：系統的開發不只是需要公共衛生人員、衛生官員、醫生等的投入，更需要資訊及電腦應用人才的配合，因此美國政府大量啟用與培養資訊人才，進而發展使用工具。

四、生物統計雖已有很好的基礎應用，但利用於 syndromic surveillance system 仍只是於起步階段。Real time、automatic 結合 G. I. S. Alert 是 syndromic surveillance 發展的主要趨勢。

五、病患資料的隱私相當受保護及重視。

六、國內在 Syndromic surveillance 發展上遭遇的問題，在美國仍相同存在，如醫師 free text 的病歷使用習慣，terminology 的不一致，HL7 在醫院的落實推廣、醫院在資料分享的配合度受 Fonding 的左右等。

七、大規模有系統性的進行系統發展，如：RODS、NRDM、ESSENS，目前進行相關的系統則有：

1. 英國利用 Call center 的 health advise service 統計相關的症狀分布，目前通報最多為：

（1）腸胃炎（嘔吐、腹瀉）

（2）類流行性感冒（發燒、咳嗽等）

（3）紅疹

（4）呼吸困難

以了解是否有地區性的疫情。

2. 紐約州監視藥局的銷售情形，分類各式藥品所適用的症狀，例如：退燒藥物的銷售情形，得以早期得知症狀的流行（NRDM）。

3. RODS 以 VPN 的網路連線，結合 HL7 的交換標準，即時監控。如：紐約州 911 電話通報中心及各大醫院急診室的病患就醫情況，結合後端 Beyien 的統計分析各症狀的分布，以提早得知是否有受生物恐怖攻擊或其他疾病聚集發生。

八、紐約州症狀監視系統，自 2000 年開始進行監視，最初有 6 家醫院加入，以生物戰之經費，進行生

物戰有關之疾病症狀監視，以急診室及實驗室為通報對象。發展 11 群診斷/症狀通報項目。包含：肺炎、ARDS、發燒/紅疹及疾病群聚事件監視等。

九、加州症狀監視系統，自 2001 年 10 月 1 日開始，起因為監視生物戰因素，以 12 家醫院急診室篩檢站所蒐集之資料、911 電話中心及動物感染傳染病通報。包含：

1. 類流行性感冒
2. 發燒/意識狀態改變
3. 發燒/紅疹
4. 腹瀉/脫水
5. 吞嚥困難/意識狀態改變
6. ARDS（呼吸困難）

911 電話中心統計相關的症狀分布，目前通報最多為：發燒、咳嗽、紅疹、腸胃炎、抽搐、呼吸困難等。

## 肆、建議

自國外遭受生物恐怖攻擊及國內 SARS 疫情爆發後，監視系統之重要性，實不容忽視，但除了即時監控疫情之外，建立預警值，提早提出警訊外，能於疫情初期即早發現並立即監控疫情，實為目前所最迫切的。非診斷性之即時監視系統中，症狀監視系統是較符合目前國內能先行試辦之系統，我國建立「症狀監視系統」之建議如下：

一、國內在症狀通報上比美國起步較晚，主要原因在於缺乏基礎的 Health information 作為 basal line，如：Test Normalization 及 Natural language process 在病例上均呈 Free text 格式，在國內極少有此方面的研究，而這些都是電子病歷的基礎。

二、目前國內仍停留於被動式的通報方式，應加入主動監視系統以提高發現聚集病例的時效性。

三、Fonding 仍是決定成功與否的主要因素，美國為建立 Public Health information Network ，於 2002 年即投入壹百萬美金，讓

衛生界及學界建立系統評估，國內於此部分資金的投入非常有限，仍有待適度的提昇。

四、國內的發展未能有系統性的規劃與投入，仍只是於科技計畫的研究，也未能有效的整合及持續性評估，建議應成立一組工作團隊，包含流病人員（40% ）、生物統計人員（20% ）、Health information 人員（30% ）、及醫師（10% ）以持續系統建立與評估。

五、在 IC 卡的推動下，本局可以與健保局合作，利用每日上傳的病歷及用藥資料，應用於各式的監測，如可利用 ICD-9 或病人主訴症狀結合 Zip code 來監視疾病的聚集，雖然不是 Real time，但仍能做到近乎 Real time 的多元化監視。

六、台灣有署立醫院的架構，可用來建立 Real time 的自動化監測。建議初期以計畫施行，先選擇 3 至 4 家醫院試辦，試辦醫院的急診資訊系統為資料輸入端，本局資訊中心建置主機系統接收資料為資料接收及儲存端，並建立資料分析

模式，待系統穩定後再行擴大辦理。

七、未來訂定症狀項目，可參考本局傳染病監視系統監視疾病項目，以為未來之比較，另症狀監視系統可監視某一症狀之流行，但通報資料如有異常，仍需配合疫調及檢驗，否則仍無法得知，係何種疾病流行。

八、如為完成上述功能，有必要建立國家健康資料庫，以pile line mult-stream的架構，建構台灣健康促進與提昇的新世紀。

# 附件一

# 2003 National Syndromic Surveillance Conference
# Brief Agenda

## Thursday, October 23rd

| | |
|---|---|
| **8:30-10:00** | **Introduction, Overview and Context** |
| 10:00 - 10:30 | Break |
| **10:30 - 12:00** | **Findings from Model Systems** |
| 12:00 - 1:30 | Lunch and Funding Opportunities |
| **1:30 - 3:00** | **National Resources under Development** |
| 3:00 - 3:30 | Break |
| **3:30 - 5:00** | **Have Syndromic Surveillance Systems Been Useful?** |
| **5:00 - 7:00** | **Vendor and Poster Sessions - Cocktail Party** |

## Friday, October 24th

There are two concurrent sessions on Friday. Track A will meet in room 20 on the second floor, with a target audience of those interested in research and development. Track B will meet in Hosack Hall with a target audience of public health practitioners.

| | Track A. Research Methods | Track B. Public Health Practice |
|---|---|---|
| 8:00 - 9:30 | **Aberration Detection**<br>**Temporal and Spatial-Temporal Methods** | **Local & State Health Departments: Experiences and Challenges** |
| 9:30 - 11:00 | **Adjustment for Natural Variation** | **Managing Relationships with Data Providers** |
| 11:00 - 11:20 | Break | Break |
| 11:20 - 12:30 | **Syndrome Definitions/ Syndrome Groupings** | **Legal Perspectives/HIPPA.** |
| 12:30 - 2:00 | Lunch | Lunch |
| 2:00 - 3:30 | **Outbreak Simulations for Performance Testing** | **Investigation of Signals** |
| 3:30 - 4:00 | Break | Break |

| | |
|---|---|
| **4:00 - 4:30** | **Lessons Learned- Research Methods Section** |
| **4:30 - 5:00** | **Lessons Learned – Public Health Practice** |
| **5:00 - 5:30** | **Closing** |

# 2003 National Syndromic Surveillance Conference
# Detailed Agenda for Thursday, October 23

### Registration and Breakfast

7:30 – 9:00 Registration in the first floor lobby. Breakfast will be served on the second floor

### Session 1
### Introduction, Overview and Context

8:30 - 10:00

*Introduction and Welcoming Address*
**Farzad Mostashari** – NYC DOHMH
**Dan Sosin** – CDC

*What is Syndromic Surveillance?*
**Kelly Henning** – NYC DOHMH

*The Challenge*
**Seth Foldy** – The City of Milwaukee Health Department

*Evaluation Challenges*
**Dan Sosin** – CDC

### Coffee Break

10:00 - 10:30 Coffee and refreshments are available on the first and second floors

### Session 2
### Findings from Model Systems
### Moderated by Patrick Kelley, The National Academies Institute of Medicine

10:30 - 12:00

*Syndromic Surveillance in New York City*
**Rick Heffernan** – NYC DOHMH

*The Real-Time Outbreak and Disease Surveillance (RODS) System*
**Michael Wagner** – University of Pittsburgh

*National Capital Region Collaborative Disease Surveillance Network Using ESSENSE*
**Joseph Lombardo** – Johns Hopkins University

### Lunch and Funding Opportunities Session

12:00 - 1:30 Lunch service on the second floor and in Room 440. You might also take your lunch in the boxes provided and sit in Central Park. Overflow seating, should the weather not cooperate, will be available in the balcony of Hosack Hall (accessible via second floor).

In Room 20 there will be a discussion of funding opportunities for syndromic surveillance and mechanisms to obtain funding.

## Session 3
## National Resources Under Development – Moderated by John Loonsk, CDC

1:30 - 3:00 *The Biosense Initiative and National Resources*
**John Loonsk** - Centers for Disease Control

*The National Bioterrorism Syndromic Surveillance Demonstration Program*
**Richard Platt** – Harvard Medical School

*National Retail Data Monitoring System*
**Michael Wagner** – University of Pittsburgh

## Coffee Break

3:00 - 3:30 Coffee and refreshments are available on the first and second floors

## Session 4
## Have Syndromic Surveillance Systems Been Useful?
## Moderated by Julie Pavlin, WRAIR

### Outbreak Detection

3:30 - 5:00 *Investigation of Diarrheal Illness Detected through Syndromic Surveillance Afte a Massive Power Outage, New York City, August 2003*
**Melissa Marx** – NYC DOHMH

*A National Symptom Surveillance System in the UK Using Calls to a Telephone Health Advice Service*
**Duncan Cooper** – Heartlands Hospital; Birmingham, UK

### Other Uses

*Use of Pharmacy Data to Evaluate Impact of Smoking Regulations in Sales of Nicotine Replacement Therapies in New York City*
**Kristi Metzger** – NYC Department of Health and Mental Hygiene

*Conducting Population Behavioral Health Surveillance suing Automated Diagnostic and Pharmacy Data Systems*
**Julie Pavlin** – WRAIR

## Poster and Vendor Session with Reception

5:00 - 7:00 Please join us for a poster session, where 50 posters will display research from around the world. Concurrently, six vendors will be displaying their products in Reception Room 1. Wine and beer will be served.

# 2003 National Syndromic Surveillance Conference
# Detailed Agenda for Friday, October 24
# Track A – Research Methods
# Track A will be meeting in Room 20

## Session 5.A
## Aberration Detection: Temporal and Spatial-Temporal Methods
## Moderated by David Madigan, Rutgers University

8:00 – 9:30

*Bayesian Approaches to Syndromic Surveillance*
**Andrew Lawson** – University of South Carolina

*Syndromic Surveillance for Post-Vaccination Adverse Events Using SPRT: A Retrospective Analysis*
**Margarette Kolczak** – CDC NIP

*The Role of Data Aggregation in Syndromic Surveillance with Applications in ESSENCE II*
**Howard Burkom** – Johns Hopkins University

## Session 6.A
## Adjustment for Natural Variation – Moderated by Henry Rolka, CDC

9:30 – 11:00

*An Integrated Algorithm for a Statistical Detection of Peaks: The Syndromic surveillance for the Athens 2004 Olympics*
**Urania Dafni** – University of Athens

*WSARE v3.0: Accounting for a Changing Baseline*
**Weng-Keen Wong** – Carnegie Mellon University

*Syndromic Surveillance without Denominator Data: The Space-Time Permutation Scan Statistic*
**Martin Kulldorff** – Harvard Pilgrim Healthcare

## Coffee Break

11:00 – 11:30 Coffee and refreshments are available on the first and second floors

## Session 7.A
## Syndrome Definitions/Syndrome Groupings
## Moderated by Farzad Mostashari, NYC DOHMH

11:30 – 12:30

*A Comparison of Two Major Emergency Department Free-Text Chief Complaint Coding Systems*
**Christina Mikocz** – Rush University Med. Ctr.

*Syndrome Definitions for Disease Associated with Critical Agents*
**Virginia Foster** – WRAIR

*Normalization of Free Text for Syndromic Surveillance*
**Alan Shapiro** – NYU School of Medicine

### Lunch

12:30 - 2:00 Lunch service on the second floor and in Room 440. You may also take your lunch in the boxes provided and sit in Central Park. Overflow seating, should the weather not cooperate, will be available in the balcony of Hosack Hall (accessible via second floor).

### Session 8.A
### Outbreak Simulations for Performance Testing
### Moderated by Martin Kulldorff, Harvard Pilgrim Healthcare

2:00 – 3:30

*Statistical Measures for Evaluation of Methods for Syndromic Surveillance*
**Marianne Frisén** – University of Gothenburg

*Syndromic Surveillance In Vitro: Outbreak Detection Performance using Limited Feature Set Simulation*
**Kenneth Mandl** – Harvard Medical School

*Evaluating Biosurveillance Prototypes*
**David Siegrist** – Potomac Institute for Policy Studies

### Coffee Break

3:30 – 4:00 Coffee and refreshments are available on the first and second floors

# 2003 National Syndromic Surveillance Conference
# Detailed Agenda for Friday, October 24
# Track B – Public Health Practice
# Track B will be meeting in Hosack Hall

## Session 5.B
## Local & State Health Departments: Experiences and Challenges
## Moderated by Christine Hahn, Idaho State DOH and the CSTE

8:00 – 9:30

*Fishing for Sharks in a Rowboat: Developing, Using and Maintaining a Sophisticated Surveillance System in Bergen County, NJ*
**Marc Palladini** – Bergen County, NJ DOH

*2003 Update for Connecticut Hospital Admissions Syndromic Surveillance (HASS)*
**Zygmunt Dembek** – Connecticut DOH

*Syndromic Surveillance for Bioterrorism in Santa Clara County, CA: October 1,2001 to September 30, 2003*
**Mujib Rahman** – County of Santa Clara, CA DOH

## Session 6.B
## Managing Relationships with Data Providers
## Moderated by Richard Platt, Harvard Medical School

9:30 – 11:00

*Managed Care Organizations as Data Providers for National Syndromic Surveillance Systems: Motivation and Concerns*
**Andrew Nelson** – Health Partners, Research Foundation

*Data Provider Relationships: Pros, Cons and Considerations*
**Andrew Kress** – Surveillance Data, Inc.

*Emergency Department Data Providers*
**Dennis Cochrane** – Emergency Medical Assoc.-NJ

## Coffee Break

11:00 – 11:30 Coffee and refreshments are available on the first and second floors

## Session 7.B
## Legal Perspectives/HIPPA
## Moderated by James Hodge, Johns Hopkins School of Public Health

11:30 – 12:30

*Legal Perspective/HIPAA*
**James Hodge** – Johns Hopkins School of Public Health

*Experience of Syndromic Surveillance Systems with issues of Patient Privacy and HIPAA*
**James Gibson** – South Carolina DOH

*Legal Perspectives of Implementing Syndromic Surveillance Systems – Utah Case Study*
**Per Gesteland** – Intermountain Health Care, University of Utah

### Lunch

12:30 - 2:00 Lunch service on the second floor and in Room 440. You may also take your lunch in the boxes provided and sit in Central Park. Overflow seating, should the weather not cooperate, will be available in the balcony of Hosack Hall (accessible via second floor).

### Session 8.B

## Investigation of Signals

## Moderated by Don Weiss, NYC DOHMH

2:00 – 3:30 *Field Investigations of Emergency Department Syndromic Surveillance Signals, New York City*
**Linda Steiner-Sichel** – NYC DOHMH

*Should We Be Worried? Investigation of Signals Generated by an Electronic Syndromic Surveillance System*
**William Terry** – Westchester County, NY DOH

*B-Safer Medical Surveillance Program*
**Edith Umland** – New Mexico Health Sciences

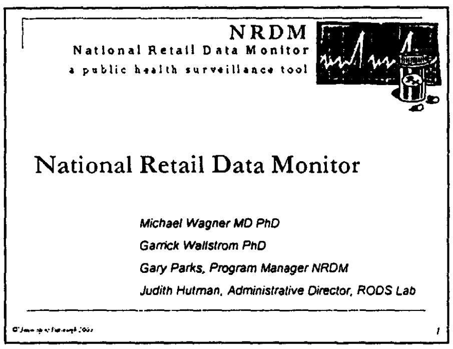
### Coffee Break

3:30 – 4:00 Coffee and refreshments are available on the first and second floors

## Session 9
## Conclusions – Hosack Hall

| | |
|---|---|
| 4:00 - 4:30 | *Lessons Learned - Research Methods Section*<br>**Farzad Mostashari** - NYC DOHMH |
| 4:30 – 5:00 | *Lessons Learned – Public Health Practice*<br>**Daniel Sosin** - CDC |
| 5:00 – 5:30 | *Closing Observations--Syndromic Surveillance: Lessons Learned and Unanswered Questions*<br>**James W. Buehler** - Emory University |

# 附件二

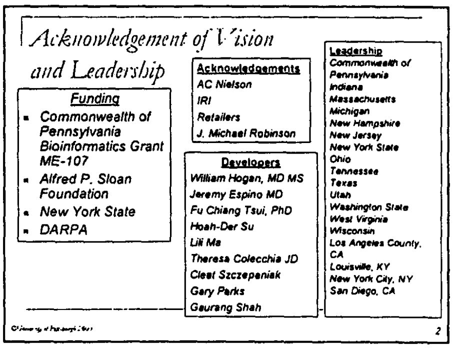


NRDM
National Retail Data Monitor
a public health surveillance tool

# National Retail Data Monitor

Michael Wagner MD PhD
Garrick Wallstrom PhD
Gary Parks, Program Manager NRDM
Judith Hutman, Administrative Director, RODS Lab

1



## Acknowledgement of Vision and Leadership

**Funding**

- Commonwealth of Pennsylvania Bioinformatics Grant ME-107
- Alfred P. Sloan Foundation
- New York State
- DARPA

**Acknowledgements**

AC Nielson
IRI
Retailers
J. Michael Robinson

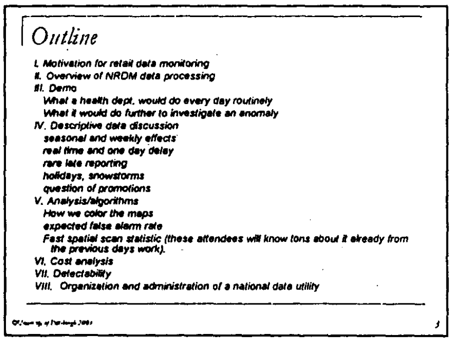
**Developers**

William Hogan, MD MS
Jeremy Espino MD
Fu Chiang Tsui, PhD
Hoah-Der Su
Lili Ma
Theresa Colecchia JD
Cleat Szczepaniak
Gary Parks
Gaurang Shah

**Leadership**

Commonwealth of Pennsylvania
Indiana
Massachusetts
Michigan
New Hampshire
New Jersey
New York State
Ohio
Tennessee
Texas
Utah
Washington State
West Virginia
Wisconsin
Los Angeles County, CA
Louisville, KY
New York City, NY
San Diego, CA

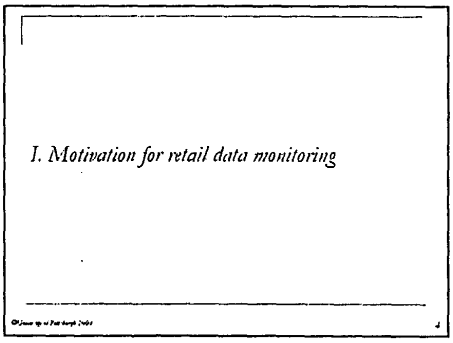
2



## Outline

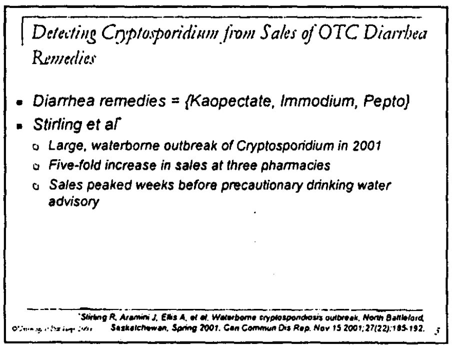
I. Motivation for retail data monitoring
II. Overview of NRDM data processing
III. Demo
What a health dept. would do every day routinely
What it would do further to investigate an anomaly
IV. Descriptive data discussion
seasonal and weekly effects
real time and one day delay
rare late reporting
holidays, snowstorms
question of promotions
V. Analysis/algorithms
How we color the maps
expected false alarm rate
Fast spatial scan statistic (these attendees will know tons about it already from the previous days work).
VI. Cost analysis
VII. Detectability
VIII. Organization and administration of a national data utility

3



## I. Motivation for retail data monitoring

4

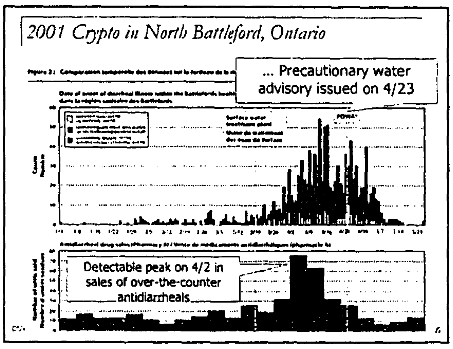


## Detecting Cryptosporidium from Sales of OTC Diarrhea Remedies

- Diarrhea remedies = {Kaopectate, Immodium, Pepto}
- Stirling et al[*]
  - Large, waterborne outbreak of Cryptosporidium in 2001
  - Five-fold increase in sales at three pharmacies
  - Sales peaked weeks before precautionary drinking water advisory

[*] Stirling R, Aramini J, Ellis A, et al. Waterborne cryptospondiosis outbreak, North Battleford, Saskatchewan, Spring 2001. Can Commun Dis Rep. Nov 15 2001;27(22):185-192.

5

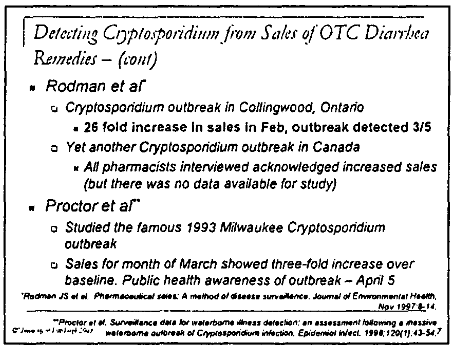


## 2001 Crypto in North Battleford, Ontario

... Precautionary water advisory issued on 4/23

Detectable peak on 4/2 in sales of over-the-counter antidiarrheals

6

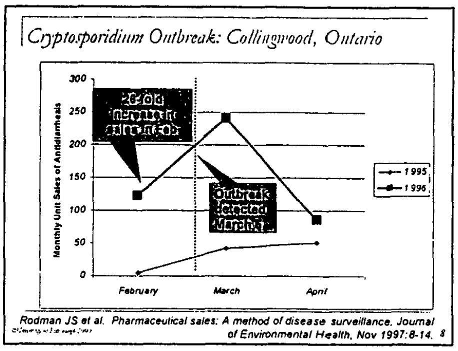


## Detecting Cryptosporidium from Sales of OTC Diarrhea Remedies – (cont)

- Rodman et al*
  - Cryptosporidium outbreak in Collingwood, Ontario
    - **26 fold increase in sales in Feb, outbreak detected 3/5**
  - Yet another Cryptosporidium outbreak in Canada
    - All pharmacists interviewed acknowledged increased sales (but there was no data available for study)
- Proctor et al**
  - Studied the famous 1993 Milwaukee Cryptosporidium outbreak
  - Sales for month of March showed three-fold increase over baseline. Public health awareness of outbreak – April 5

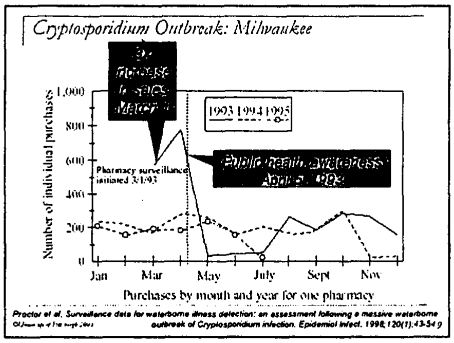
*Rodman JS et al. Pharmaceutical sales: A method of disease surveillance. Journal of Environmental Health, Nov 1997:8-14.

**Proctor et al. Surveillance data for waterborne illness detection: an assessment following a massive waterborne outbreak of Cryptosporidium infection. Epidemiol Infect. 1998;120(1):43-54.



## Cryptosporidium Outbreak: Collingwood, Ontario

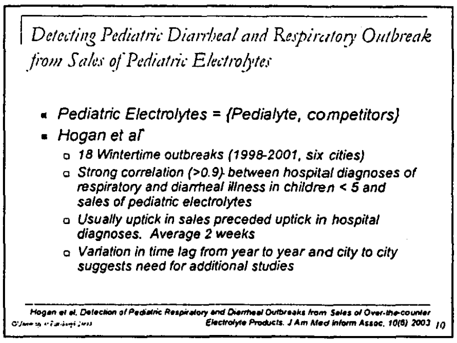
Rodman JS et al. Pharmaceutical sales: A method of disease surveillance. Journal of Environmental Health, Nov 1997:8-14.



## Cryptosporidium Outbreak: Milwaukee

Purchases by month and year for one pharmacy

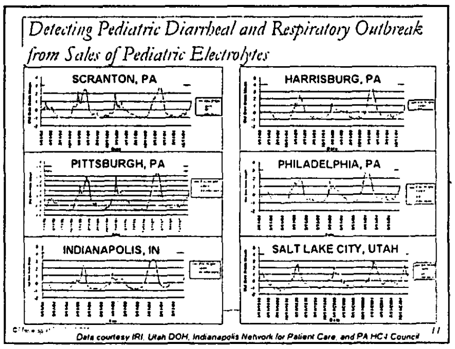
Proctor et al. Surveillance data for waterborne illness detection: an assessment following a massive waterborne outbreak of Cryptosporidium infection. Epidemiol Infect. 1998;120(1):43-54



## Detecting Pediatric Diarrheal and Respiratory Outbreak from Sales of Pediatric Electrolytes

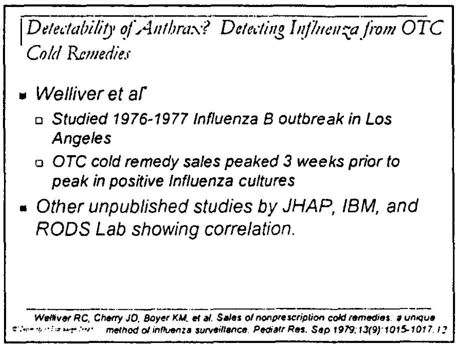
- Pediatric Electrolytes = {Pedialyte, competitors}
- Hogan et al*
  - 18 Wintertime outbreaks (1998-2001, six cities)
  - Strong correlation (>0.9) between hospital diagnoses of respiratory and diarrheal illness in children < 5 and sales of pediatric electrolytes
  - Usually uptick in sales preceded uptick in hospital diagnoses. Average 2 weeks
  - Variation in time lag from year to year and city to city suggests need for additional studies

*Hogan et al. Detection of Pediatric Respiratory and Diarrheal Outbreaks from Sales of Over-the-counter Electrolyte Products. J Am Med Inform Assoc. 10(6) 2003



## Detecting Pediatric Diarrheal and Respiratory Outbreak from Sales of Pediatric Electrolytes

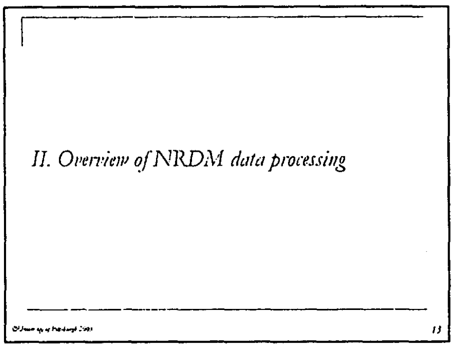
Data courtesy IRI, Utah DOH, Indianapolis Network for Patient Care, and PA HC4 Council



## Detectability of Anthrax? Detecting Influenza from OTC Cold Remedies

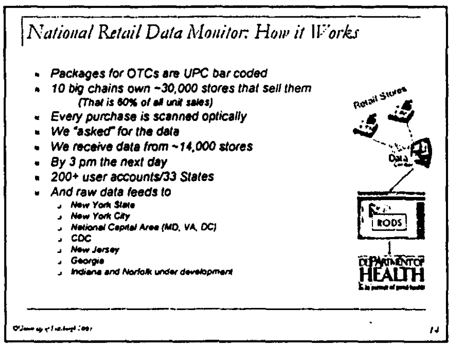
- Welliver et al*
  - Studied 1976-1977 Influenza B outbreak in Los Angeles
  - OTC cold remedy sales peaked 3 weeks prior to peak in positive Influenza cultures
- Other unpublished studies by JHAP, IBM, and RODS Lab showing correlation.

*Welliver RC, Cherry JD, Boyer KM, et al. Sales of nonprescription cold remedies: a unique method of influenza surveillance. Pediatr Res. Sep 1979;13(9):1015-1017.



## II. Overview of NRDM data processing

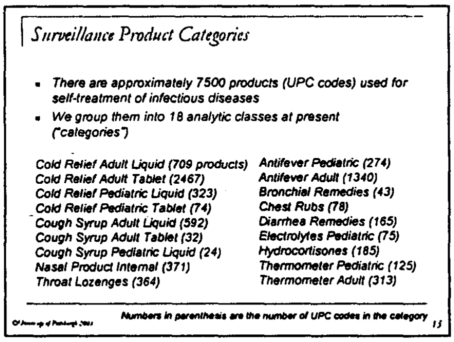
©University of Pittsburgh 2003 13



## National Retail Data Monitor: How it Works

- Packages for OTCs are UPC bar coded
- 10 big chains own ~30,000 stores that sell them
  (That is 60% of all unit sales)
- Every purchase is scanned optically
- We "asked" for the data
- We receive data from ~14,000 stores
- By 3 pm the next day
- 200+ user accounts/33 States
- And raw data feeds to
  - New York State
  - New York City
  - National Capital Area (MD, VA, DC)
  - CDC
  - New Jersey
  - Georgia
  - Indiana and Norfolk under development

©University of Pittsburgh 2003 14



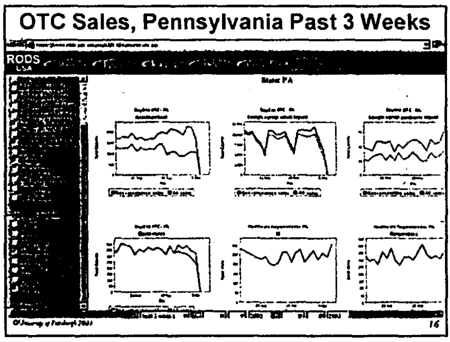
## Surveillance Product Categories

- There are approximately 7500 products (UPC codes) used for self-treatment of infectious diseases
- We group them into 18 analytic classes at present ("categories")

| | |
|---|---|
| Cold Relief Adult Liquid (709 products) | Antifever Pediatric (274) |
| Cold Relief Adult Tablet (2467) | Antifever Adult (1340) |
| Cold Relief Pediatric Liquid (323) | Bronchial Remedies (43) |
| Cold Relief Pediatric Tablet (74) | Chest Rubs (78) |
| Cough Syrup Adult Liquid (592) | Diarrhea Remedies (165) |
| Cough Syrup Adult Tablet (32) | Electrolytes Pediatric (75) |
| Cough Syrup Pediatric Liquid (24) | Hydrocortisones (185) |
| Nasal Product Internal (371) | Thermometer Pediatric (125) |
| Throat Lozenges (364) | Thermometer Adult (313) |

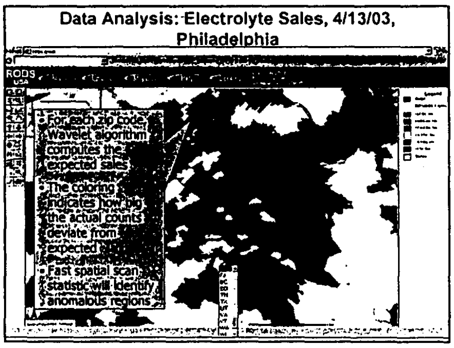
Numbers in parenthesis are the number of UPC codes in the category

©University of Pittsburgh 2003 15



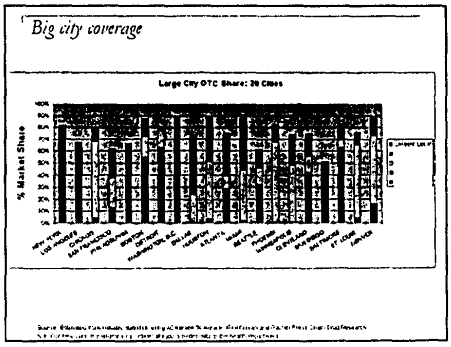
## OTC Sales, Pennsylvania Past 3 Weeks

©University of Pittsburgh 2003 16



## Data Analysis: Electrolyte Sales, 4/13/03, Philadelphia

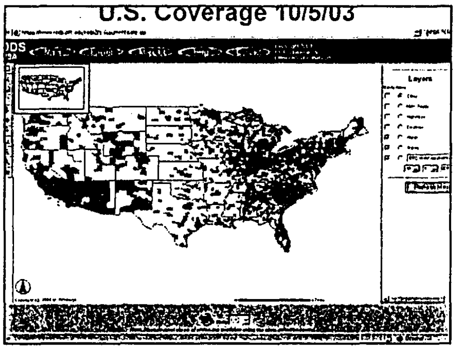
- For each zip code, Wavelet algorithm computes the expected sales
- The coloring indicates how big the actual counts deviate from expected [illegible]
- Fast spatial scan statistic will identify anomalous regions



## Big city coverage

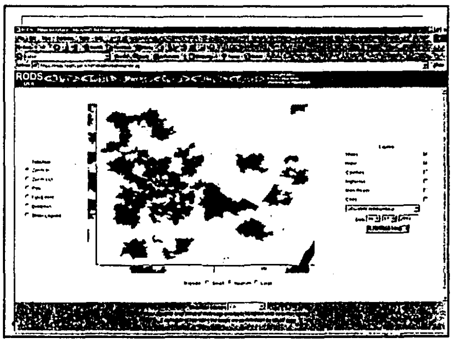
Source: [illegible]

## U.S. Coverage 10/5/03

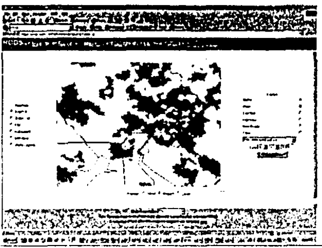


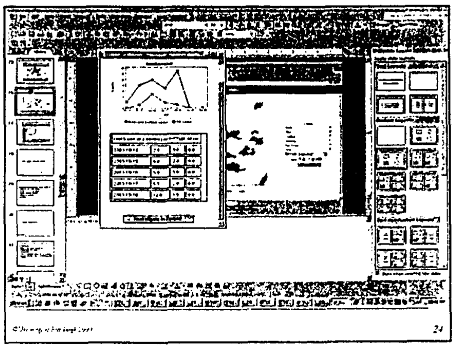
## III. Demo

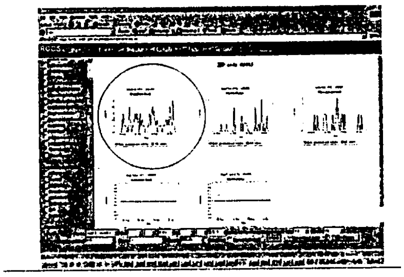
Real-time Outbreak and Disease Surveillance
Username
Password



## *Antidiarrheals, Philadelphia 10/17/03*







## IV. Descriptive data discussion

## Qualitative Description

- Real time and one day delay
- Promotions
- Adding a retailer
- Seasonal, Weekly and Day-of-week effects
- Autoregressive effects
- Holidays
- Explainable anomalies (snowstorms, bulk purchases)

## Category Maintenance

- UPC codes can change
- There is a master that retailers can download
- The master must be maintained

## V. Analysis/algorithms

## Coloring of the Maps

- Map is colored based on the number of standard deviations above the predicted sales
  - Green: < 0.5 standard deviations
  - Blue: 0.5 – 2.0 standard deviations
  - Yellow: 2 – 3 standard deviations
  - Orange: 3 – 4 standard deviations
  - Red: > 4 standard deviations
- Wavelet-based predictions account for seasonal, weekly and weekend effects.

## Expected Number of False Alarms

- Currently have data from 7500 zip codes nationally.
- Each zip code is expected to appear Red 0.0365 times per year when there is no outbreak.
- 274 red zip codes expected per year nationally
- Later we discuss cost of such investigations

## Fast Spatial Scan Statistic

- You have heard in depth about spatial scan statistic
- We run an efficient version developed by Dan Pelleg and Andrew Moore that can process the entire countries retail data
- Rationale: although the cover of New Yorkers magazine suggests that a vast desert begins a few miles west of the Hudson, we did some research suggesting otherwise.
- Outbreaks don't subscribe to the New Yorker framework

## List of other Analytic Issues that Will Keep us All Employed for Some Time

- What are the optimal categories?
- How to handle late reports?
- Normalization?
- What are the confounders and how to model them (day of week, sales, snowstorms, power outages)?
- How to integrate with other types of data that have different temporal properties?

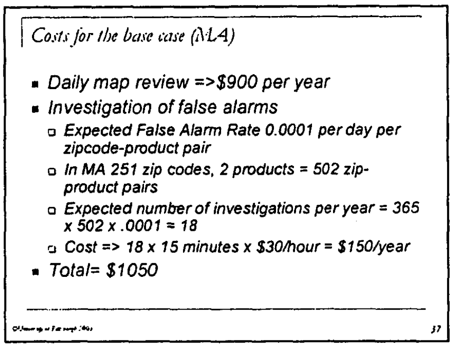
## VI. Cost analysis

## Cost-Benefit Tradeoff

- Benefit: Does the NRDM provide information of value, such as earlier indication of the existence of an outbreak of disease
- Cost: What is the cost at which this potential benefit may be obtained, measured in terms of both economic cost of operating the system as well as the cost of investigating anomalies in sales detected by the system?
- Finally, what policy is optimal (e.g., choice of detection threshold and actions to be taken at that threshold)

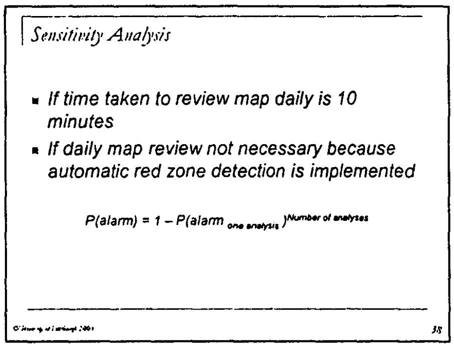
## Base Case

- The NRDM maps are reviewed daily in Massachusetts to track diarrheal remedies and pediatric electrolytes
- Daily review takes approximately 5 minutes.
- Only red zip codes are investigated further
- Further investigation of red zip code takes approximately 15 minutes



## Costs for the base case (MA)

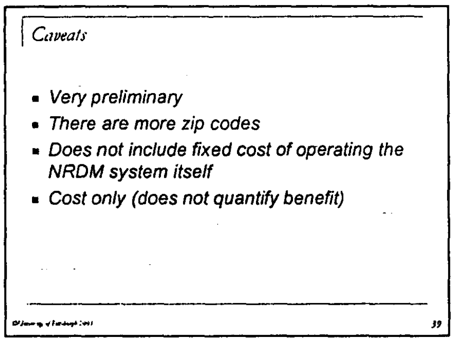
- Daily map review =>$900 per year
- Investigation of false alarms
  - Expected False Alarm Rate 0.0001 per day per zipcode-product pair
  - In MA 251 zip codes, 2 products = 502 zip-product pairs
  - Expected number of investigations per year = 365 x 502 x .0001 = 18
  - Cost => 18 x 15 minutes x $30/hour = $150/year
- Total= $1050



## Sensitivity Analysis

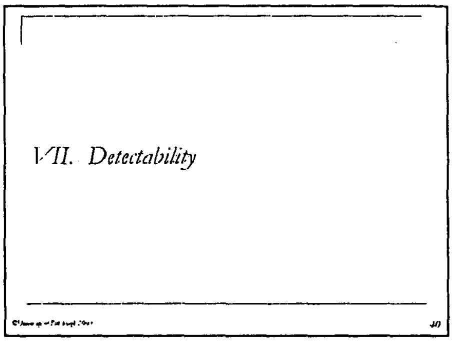
- If time taken to review map daily is 10 minutes
- If daily map review not necessary because automatic red zone detection is implemented

$$P(\text{alarm}) = 1 - P(\text{alarm}_{\text{one analysis}})^{\text{Number of analyses}}$$

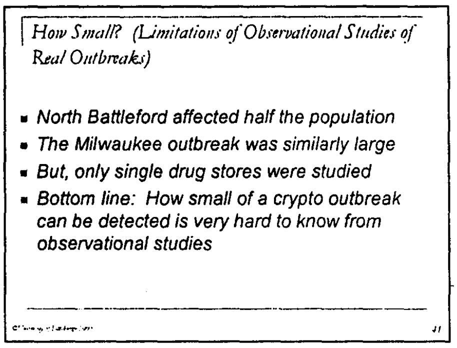


## Caveats

- Very preliminary
- There are more zip codes
- Does not include fixed cost of operating the NRDM system itself
- Cost only (does not quantify benefit)

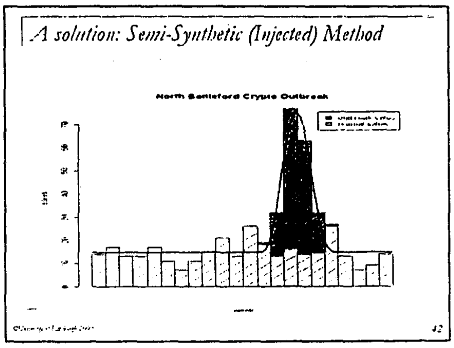


## VII. Detectability



## How Small? (Limitations of Observational Studies of Real Outbreaks)

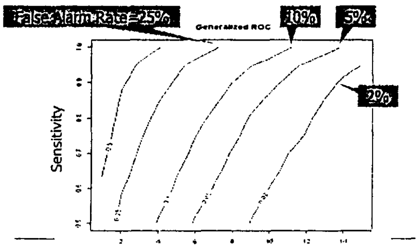
- North Battleford affected half the population
- The Milwaukee outbreak was similarly large
- But, only single drug stores were studied
- Bottom line: How small of a crypto outbreak can be detected is very hard to know from observational studies



## A solution: Semi-Synthetic (Injected) Method

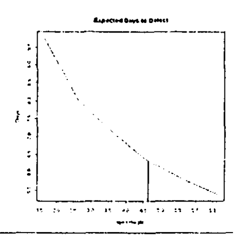
## *Detectability of a Small Spike Injected into Zip Code Data*

*(repeated for every day of 2001 and averaged over 500 zip codes)*



Detection Delay (Days from Start of Inject)

## *Expected Days to Detection*



## *VIII. Organization and administration of a national data utility*

## *Organization & Administration*

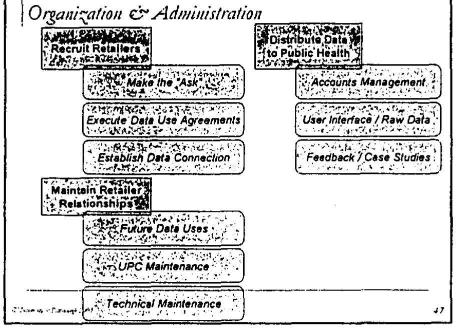
**Working Group**
*Joel Hersh, PA*
*Bob Campbell, OH*
*Stephanie Miller, NH*
*Michael Colette, GA*

DEPARTMENT OF HEALTH
... in pursuit of good health

**Monthly Phone Meeting Participants**
*All users*

**Helping to Make the Ask**
*Pennsylvania*
*Florida*
*Idaho*
*Indiana*
*Massachusetts*
*Michigan*
*Montana*
*New Hampshire*
*New Jersey*
*New York State*
*Ohio*
*Tennessee*
*Texas*
*Utah*
*Washington State*
*West Virginia*
*Wisconsin*
*Los Angeles County, CA*
*Louisville, KY*
*New York City, NY*
*San Diego, CA*

## *Organization & Administration*

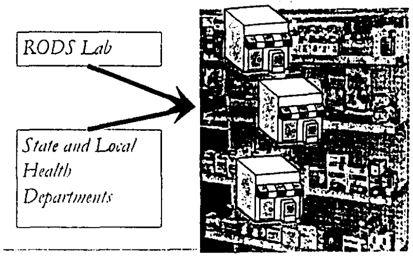


**Recruit Retailers**
- *Make the "Ask"*
- *Execute Data Use Agreements*
- *Establish Data Connection*

**Maintain Retailer Relationships**
- *Future Data Uses*
- *UPC Maintenance*
- *Technical Maintenance*

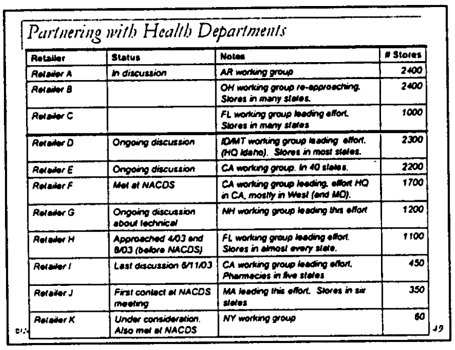
**Distribute Data to Public Health**
- *Accounts Management*
- *User Interface / Raw Data*
- *Feedback / Case Studies*

## *Making the "Ask"*



*RODS Lab*

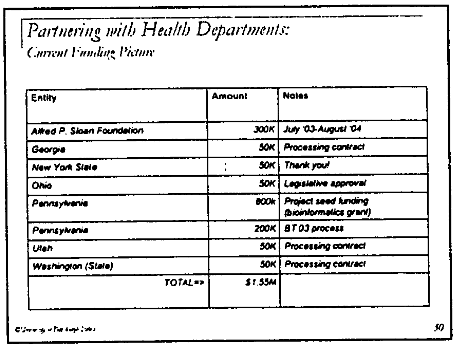
*State and Local Health Departments*



## Partnering with Health Departments

| Retailer | Status | Notes | # Stores |
|---|---|---|---|
| Retailer A | In discussion | AR working group | 2400 |
| Retailer B |  | OH working group re-approaching. Stores in many states. | 2400 |
| Retailer C |  | FL working group leading effort. Stores in many states | 1000 |
| Retailer D | Ongoing discussion | ID/MT working group leading effort. (HQ Idaho). Stores in most states. | 2300 |
| Retailer E | Ongoing discussion | CA working group. In 40 states. | 2200 |
| Retailer F | Met at NACDS | CA working group leading, effort HQ in CA, mostly in West (and MO). | 1700 |
| Retailer G | Ongoing discussion about technical | NH working group leading this effort | 1200 |
| Retailer H | Approached 4/03 and 8/03 (before NACDS) | FL working group leading effort. Stores in almost every state. | 1100 |
| Retailer I | Last discussion 6/11/03 | CA working group leading effort. Pharmacies in five states | 450 |
| Retailer J | First contact at NACDS meeting | MA leading this effort. Stores in six states | 350 |
| Retailer K | Under consideration. Also met at NACDS | NY working group | 60 |

49

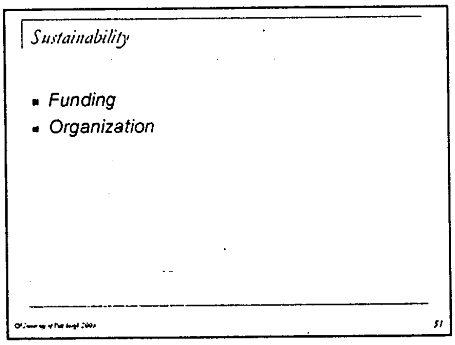


## Partnering with Health Departments:

Current Funding Picture

| Entity | Amount | Notes |
|---|---|---|
| Alfred P. Sloan Foundation | 300K | July '03-August '04 |
| Georgia | 50K | Processing contract |
| New York State | 50K | Thank you! |
| Ohio | 50K | Legislative approval |
| Pennsylvania | 800k | Project seed funding (bioinformatics grant) |
| Pennsylvania | 200K | BT 03 process |
| Utah | 50K | Processing contract |
| Washington (State) | 50K | Processing contract |
| TOTAL=> | $1.55M |  |

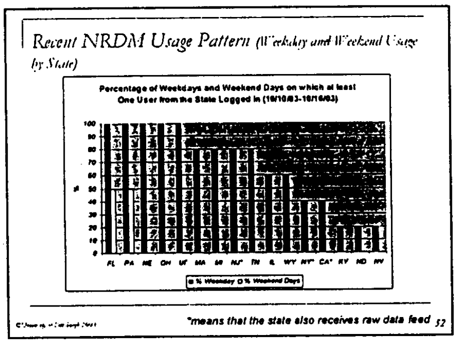
50



## Sustainability

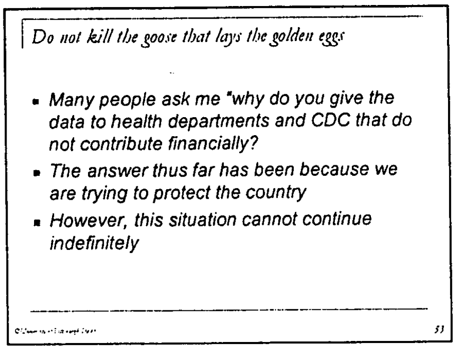
- Funding
- Organization

51



## Recent NRDM Usage Pattern (Weekday and Weekend Usage by State)

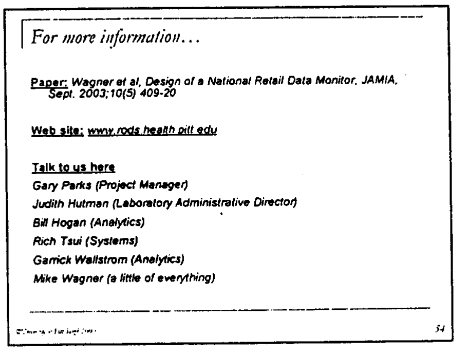
*means that the state also receives raw data feed 52



## Do not kill the goose that lays the golden eggs

- Many people ask me "why do you give the data to health departments and CDC that do not contribute financially?
- The answer thus far has been because we are trying to protect the country
- However, this situation cannot continue indefinitely

53



## For more information...

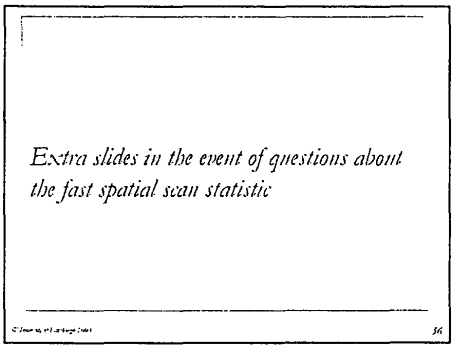
Paper: Wagner et al, Design of a National Retail Data Monitor, JAMIA, Sept. 2003;10(5) 409-20

Web site: www.rods.health.pitt.edu

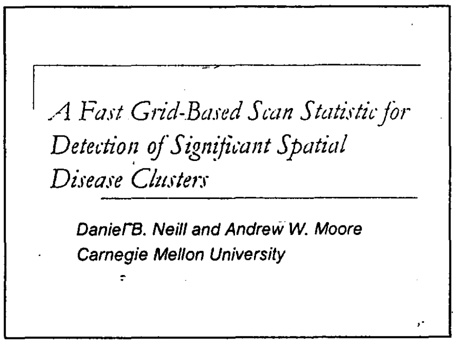
Talk to us here

Gary Parks (Project Manager)
Judith Hutman (Laboratory Administrative Director)
Bill Hogan (Analytics)
Rich Tsui (Systems)
Garrick Wallstrom (Analytics)
Mike Wagner (a little of everything)

54

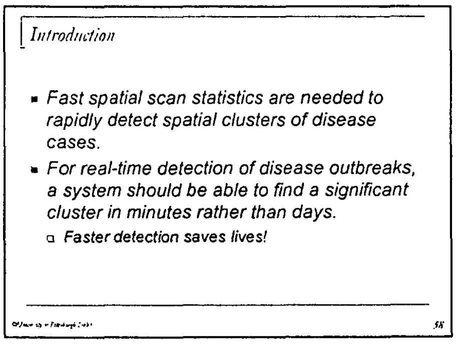


*Extra slides in the event of questions about the fast spatial scan statistic*



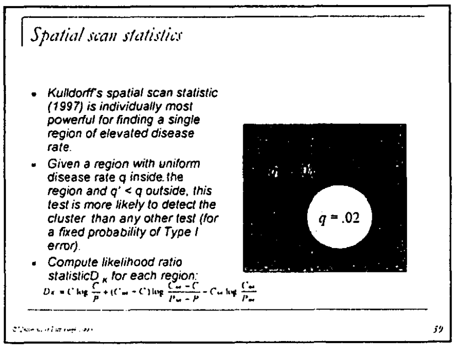
# *A Fast Grid-Based Scan Statistic for Detection of Significant Spatial Disease Clusters*

DanierB. Neill and Andrew W. Moore
Carnegie Mellon University



## *Introduction*

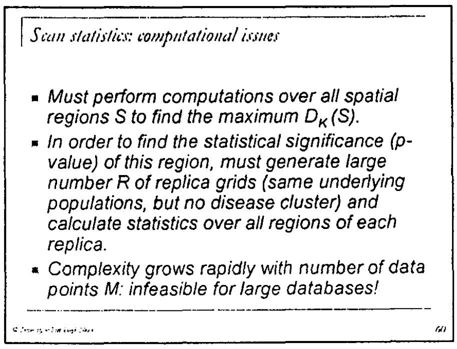
- *Fast spatial scan statistics are needed to rapidly detect spatial clusters of disease cases.*
- *For real-time detection of disease outbreaks, a system should be able to find a significant cluster in minutes rather than days.*
  - *Faster detection saves lives!*



## *Spatial scan statistics*

- *Kulldorff's spatial scan statistic (1997) is individually most powerful for finding a single region of elevated disease rate.*
- *Given a region with uniform disease rate q inside the region and q' < q outside, this test is more likely to detect the cluster than any other test (for a fixed probability of Type I error).*
- *Compute likelihood ratio statisticD $_K$ for each region:*

$$D_K = C \log \frac{C}{P} + (C_{all} - C) \log \frac{C_{all} - C}{P_{all} - P} - C_{all} \log \frac{C_{all}}{P_{all}}$$



## *Scan statistics: computational issues*

- *Must perform computations over all spatial regions S to find the maximum $D_K(S)$.*
- *In order to find the statistical significance (p-value) of this region, must generate large number R of replica grids (same underlying populations, but no disease cluster) and calculate statistics over all regions of each replica.*
- *Complexity grows rapidly with number of data points M: infeasible for large databases!*

## Grid-based scan statistics

- The standard scan statistic is slow when the number of data points M is large.
- A simple solution is to aggregate points to a uniform N x N grid: complexity is then a function of N, not M.

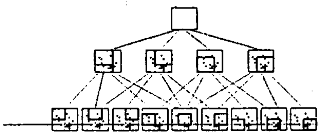
| | | | | |
|---|---|---|---|---|
| P=5000<br>C=27 | P=3500<br>C=14 | P=4500<br>C=22 | P=3000<br>C=15 | P=1000<br>C=5 |
| P=5000<br>C=26 | P=4000<br>C=17 | P=3000<br>C=12 | P=2000<br>C=12 | P=1000<br>C=4 |
| P=5000<br>C=19 | P=5000<br>C=25 | P=9000<br>C=43 | P=9000<br>C=37 | P=4000<br>C=20 |
| P=4800<br>C=18 | P=4800<br>C=20 | P=4000<br>C=40 | P=3000<br>C=22 | P=4000<br>C=16 |
| P=4700<br>C=20 | P=3000<br>C=13 | P=3000<br>C=18 | P=2000<br>C=20 | P=1000<br>C=4 |

## A naïve grid-based approach

- Search <u>all</u> square regions, return the highest value of the scan statistic, and do randomization testing.
- This is faster than the standard scan statistic when the grid is dense.
- However, still too slow for real-time detection!

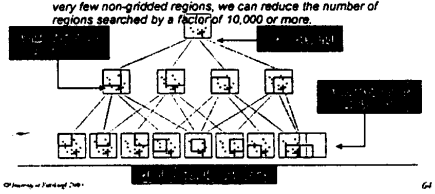
## A fast grid-based approach

- We propose a new multi-resolution algorithm which:
  - Partitions the grid into overlapping regions.
  - Performs a top-down search, first at coarse resolutions (large regions) then successively finer resolutions as necessary.
  - Prunes regions which cannot contain the maximum density region.



## "Gridded" regions

- For each resolution, we define a set of overlapping regions which cover the entire grid; only a small proportion of regions are "gridded."
- If we can confine our search to the gridded regions, and search very few non-gridded regions, we can reduce the number of regions searched by a factor of 10,000 or more.

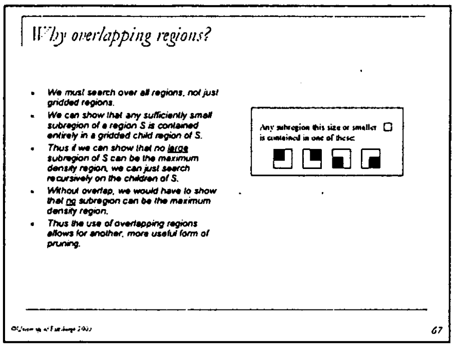


## Bounding region density

- We precompute bounds on the populations and densities of the squares contained in each gridded region.
- This can be done quickly since there are relatively few gridded regions.
- We can use this information to compute an <u>upper bound</u> on the score $D_K$ for all subregions of a given region.

## Region pruning

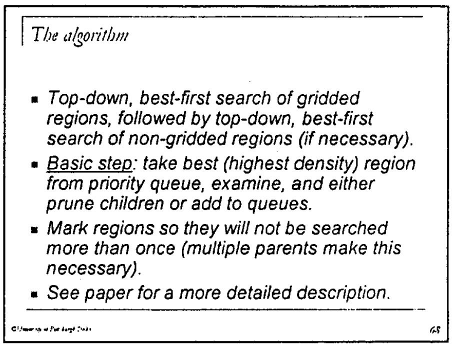
- In our top-down search, we keep track of the best region found so far, and its score.
- If the upper bound for a region is worse than the best score so far, we can prune it.
  - If no subregion can be optimal, prune completely (don't search any subregions).
  - If no large subregion can be optimal, recurse on the smaller gridded subregions.
  - If neither case applies, we must search gridded and non-gridded subregions.



## Why overlapping regions?

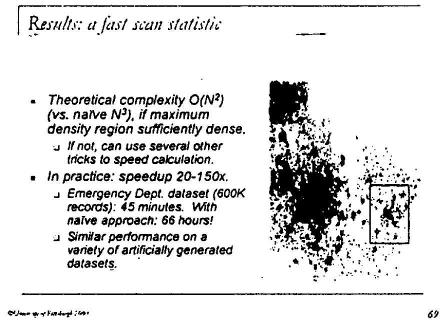
- We must search over all regions, not just gridded regions.
- We can show that any sufficiently small subregion of a region S is contained entirely in a gridded child region of S.
- Thus if we can show that no large subregion of S can be the maximum density region, we can just search recursively on the children of S.
- Without overlap, we would have to show that no subregion can be the maximum density region.
- Thus the use of overlapping regions allows for another, more useful form of pruning.

Any subregion this size or smaller □
is contained in one of these:



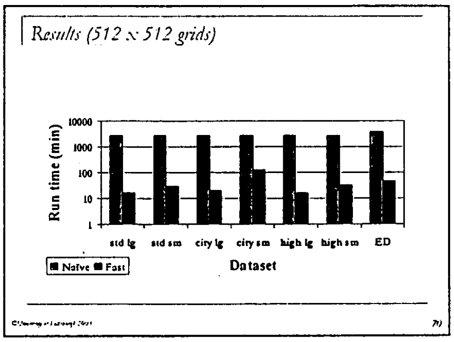
## The algorithm

- Top-down, best-first search of gridded regions, followed by top-down, best-first search of non-gridded regions (if necessary).
- Basic step: take best (highest density) region from priority queue, examine, and either prune children or add to queues.
- Mark regions so they will not be searched more than once (multiple parents make this necessary).
- See paper for a more detailed description.

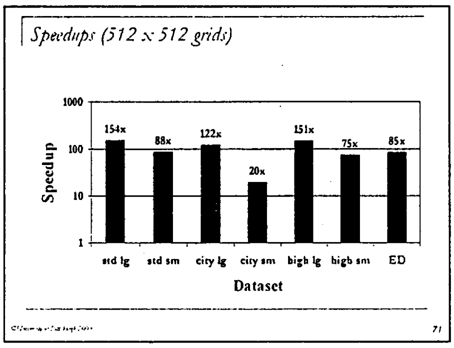


## Results: a fast scan statistic

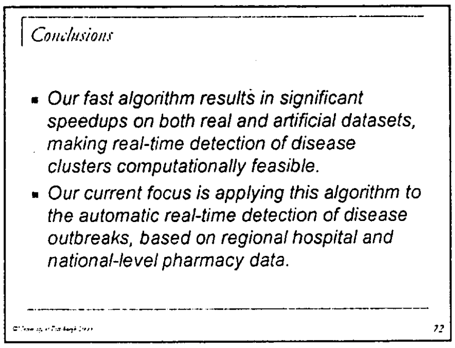
- Theoretical complexity O($N^2$) (vs. naïve $N^3$), if maximum density region sufficiently dense.
    - If not, can use several other tricks to speed calculation.
- In practice: speedup 20-150x.
    - Emergency Dept. dataset (600K records): 45 minutes. With naïve approach: 66 hours!
    - Similar performance on a variety of artificially generated datasets.



## Results (512 x 512 grids)



## Speedups (512 x 512 grids)



## Conclusions

- Our fast algorithm results in significant speedups on both real and artificial datasets, making real-time detection of disease clusters computationally feasible.
- Our current focus is applying this algorithm to the automatic real-time detection of disease outbreaks, based on regional hospital and national-level pharmacy data.

*Application of Information Technology* ■

# Design of a National Retail Data Monitor for Public Health Surveillance

Michael M. Wagner, MD, PhD, J. Michael Robinson, Fu-Chiang Tsui, PhD, Jeremy U. Espino, MD, William R. Hogan, MD

**Abstract** The National Retail Data Monitor receives data daily from 10,000 stores, including pharmacies, that sell health care products. These stores belong to national chains that process sales data centrally and utilize Universal Product Codes and scanners to collect sales information at the cash register. The high degree of retail sales data automation enables the monitor to collect information from thousands of store locations in near to real time for use in public health surveillance. The monitor provides user interfaces that display summary sales data on timelines and maps. Algorithms monitor the data automatically on a daily basis to detect unusual patterns of sales. The project provides the resulting data and analyses, free of charge, to health departments nationwide. Future plans include continued enrollment and support of health departments, developing methods to make the service financially self-supporting, and further refinement of the data collection system to reduce the time latency of data receipt and analysis.

■ **J Am Med Inform Assoc.** 2003;10:409–418. DOI 10.1197/jamia.M1357.

The rapid, early detection of disease outbreaks has become a national priority and an emerging field of research.[1] Kaufmann et al.,[2] after analyzing several bioterrorism scenarios, concluded that "delay in starting a prophylaxis program is the single most important factor leading to increased loss of life and health." The urgency of the problem is reflected in an explosion of research on new computer-based disease surveillance systems.[3–12]

Wagner et al.[1] recently discussed four possible ways to improve the earliness of outbreak detection. Prominent among them is the use of new types of surveillance data that track sales of over-the-counter (OTC) health care products such as cough syrup that are purchased early in the course of illness by sick individuals for the symptomatic treatment of illness. For common syndromes such as upper respiratory illness ("flu") and asthma, the sick are more likely to self-treat with OTC health care products than to see a physician. In a random digit dialing survey of 1,505 individuals conducted by the Consumer Healthcare Products Association in 2001, 72% (505 of 701) of those with cough, cold, "flu," or sore throat in the previous six months treated themselves with an OTC health care product. Importantly, in 42%, purchase or use of OTC preparations was their first action, and in 34%, self-observation was the first action.[13] In less than 9% was seeking professional medical care the first action. For the symptom of headache, the findings were even stronger (81% self-medication and 52% self-observation, respectively, with 4% seeking a physician as the first action).[13] In a population-based survey of 42,333 adults in the province of Ontario, Canada, only 14% of adults with upper respiratory tract infections visited a doctor, whereas 76% engaged in self-care with OTC medications.[14] Sales of OTC health care products have attractive characteristics for outbreak detection. In the United States, use of scanners and Universal Product Codes at checkout counters in retail industry stores facilitates routine collection of such data in real time. A small number of national companies own the majority of retail outlets that sell such products, and these corporations integrate their sales data at the national level in near to real time. For these reasons, the technical effort and cost to obtain these data for public health surveillance are comparatively low.

Preliminary studies suggest that sales of OTC health care products can be used for the early detection of outbreaks,[15–17] yet research progress has been slow due to the difficulty in obtaining data to adequately test the hypothesis in a sufficiently large number of sizable outbreaks. The first such study, reported in 1979, showed an association between influenza B activity and purchases of cold remedies for a single outbreak.[17] A recent study of 18 seasonal outbreaks in children showed that sales of pediatric electrolyte solutions correlated strongly with hospitalizations of children for gastrointestinal and respiratory illnesses and usually preceded the hospitalizations by more than two weeks.[15] Unpublished research shows increases in sales of OTC "flu" and cough preparations preceding increases in influenza activity as measured by outpatient billing diagnoses.[16] Other research (conducted by our laboratory and others) simulates

Affiliations of the authors: The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania (MMW, F-CT, JUE, WRH); Global Strategic Solutions, Cincinnati, Ohio (JMR).

Supported by Pennsylvania Department of Health Award number ME-01-737. The authors are indebted to Judith Hutman, Joel Hersh, William Yasnoff, James Seligman, Julie Gerberding, MD, Information Resources, Inc., ACNielsen, Procter & Gamble, and the participating retailers. The authors thank the anonymous reviewer who provided detailed comments.

Correspondence and reprints: Michael M. Wagner, MD, PhD, The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Suite 550, 100 Technology Drive, Pittsburgh, PA 15219; e-mail: <mmw@cbmi.pitt.edu>.

Received for publication: 03/06/03; accepted for publication: 05/13/03.

the effects of outbreaks of different magnitudes and time courses on sales of OTC health care products (and other types of data) to understand the smallest increase in sales that would be detectable above background levels of sales.[18,19]

A rationale for not monitoring nonspecific data such as sales of OTC health care products as a means of alerting has been presented by Broome et al.,[20] who raise concerns that data with low specificity will require an increment of hundreds of units over baseline levels before detection would occur, and that the cost of investigating false alarms may be too high.

Because of the threat of bioterrorism, the relative ease and low cost of monitoring sales of OTC health care products, and the accumulating evidence in favor of monitoring sales of OTC health care products, the Real-time Outbreak and Disease Surveillance (RODS) laboratory worked with the retail industry to build a retail data monitor for the Commonwealth of Pennsylvania. The authors quickly realized that the same effort could create a National Retail Data Monitor. This report describes the design and current status of the National Retail Data Monitor.

## Background

### Public Health Surveillance

The role of public health surveillance in general is to collect, analyze, and interpret data about biological agents, diseases, risk factors, and other health events and to provide timely dissemination of collected information to decision makers.[21–23]

Public health surveillance traditionally relies on manual operations and off-line analysis. Key components of such activities include reporting by clinicians and laboratories on notifiable diseases and dependence on astute clinicians to notice and report suspicious clusters of cases to health departments. The utility of these traditional disease surveillance approaches for the early detection of outbreaks is limited severely by delays in obtaining and analyzing the data, by the reliability of the antiquated "reportable disease" system, and by the delays that result from clinicians waiting to make reports until the diagnosis is confirmed by definitive testing.

Newer automated surveillance systems and other monitoring techniques have been developed to detect epidemics more rapidly, utilizing routinely collected and often prediagnostic data. These efforts have been pioneered by regional projects located in New York City, Washington, DC, Utah, Indiana, Seattle, Pennsylvania, and other locations.[3–11]

### Universal Product Codes

Monitoring of retail data would be impossible if not for the existence of a standard coding system for retail products. Retailers and manufacturers in North America use Universal Product Bar codes ("UPC codes") in their data systems. UPC codes are 12-digit numbers used by manufacturers to uniquely identify themselves and their products worldwide. UPC codes consist of black and white bars and a number. Retailers scan products at the cash register to detect these symbols and thereby collect data in real time about their sales. UPC codes originate from the Uniform Code Council, Inc. (UCC), a not-for-profit standards organization. The mission of the UCC is to establish and promote multi-industry standards for product identification and related electronic communication. The UCC administers the UPC codes and provides a range of standards and business solutions for over 250,000 member companies doing business in 25 major industries. A manufacturer pays an annual fee to join the council. In return, the council issues the manufacturer a six-digit *manufacturer identification number* and provides guidelines on how to use it. The manufacturer identification number is the first six digits of the UPC number, and the next five digits are the item number followed by a one-digit field used as a check digit to ensure UPC number accuracy. A person employed by the manufacturer, called the *UPC coordinator*, is responsible for assigning item numbers to products, making sure the same code is not used on more than one product, retiring codes as products are removed from the product line, and other duties. In general, every item the manufacturer sells, as well as every size package and every repackaging of the item, needs a different item code. So, a 4-ounce bottle of *Tylenol Children's Cold Great Grape Flavor—Alcohol Free, Aspirin Free,* is assigned a different item number from an 8-ounce bottle of *Tylenol Children's Cold Great Cherry Flavor—Alcohol Free, Aspirin Free.*

### Regions that Monitor OTC Sales

Surveillance systems developed by the Johns Hopkins Applied Physics Laboratory in the National Capital Area [24] and by the New York City Department of Health [25] collect and analyze sales of OTC health care products for public health surveillance. The New York City project obtains data in nine product categories from a single local chain. The National Capital Area project obtains data from two large national chains, but the data are limited to the National Capital Area. These efforts have similarities to the current project in the type of data being collected and the purpose of the systems. A key difference of the National Retail Data Monitor is that it receives nationwide data from the national data warehouses of retailers.

## Design Objectives

The purpose of the National Retail Data Monitor is to collect and analyze sales of OTC health care products to detect outbreaks of disease. Because of the relative lack of specificity of OTC health care products as indicators of disease, the types of outbreaks that the current system is intended to detect are those that involve a relatively large proportion of individuals in the geographic region being analyzed (in this case, zip code). Currently, the methods used in the National Retail Data Monitor also are focused on detecting sudden outbreaks, although this current specialization is not a fundamental limitation of the approach. In particular, the current niche for the National Retail Data Monitor is early detection of a mass exposure of a large number of people through contamination of the air, food, or water (a *cohort exposure*). Soon after such an exposure, the cohort will become symptomatic, and, depending on the symptoms, may begin self-treatment and then either recover or seek medical care. If the cohort is large enough, sales of OTC health care products will increase significantly above the normal, background level of sales.

A key requirement for early detection is to minimize the time latency between time of purchase and time when data become available for analysis. Time intervals as small as hours can make a difference when a large cohort is exposed to rapidly progressing diseases such as anthrax.[1,2,26] Therefore, a design objective is to collect and analyze the data in as near as real time as possible, with at most a day's delay from time of sale. A second requirement is completeness of sales data collection, which is important for both early detection and sensitivity to smaller outbreaks. The number of independent stores not participating in centralized data collection limits available data to less than 100% coverage of sales. The project's specific objective, therefore, is to collect sales data sufficient to achieve at least 70% market share nationally and to achieve 70% share in each of the 20 largest urban regions (Fig. 2). A third requirement concerns the need for precise spatial information in outbreak detection. Ideally, one would receive data that support spatial analysis at the level of individual store locations, or at least by the zip codes of stores.

Additional desiderata include collection of supplemental data—for example, indicating when retailers feature promotions or how day of the week affects local sales volumes. Because UPC codes change, a system for maintaining UPC code masters and mappings from UPC codes to analytic categories also is a requirement. Finally, it is essential to create an effective link between the surveillance data, public health review, and response. If the information collected by the National Retail Data Monitor is not reviewed by the intended users of the system (local, state, and federal public health authorities) and does not influence response (e.g., quarantine and medical treatment), the National Retail Data Monitor cannot have an impact on preventing morbidity and mortality and therefore will have no utility beyond supporting retrospective research.

## System Description

This section describes methods used in the National Retail Data Monitor to receive data from retailers, analyze and process the data, and define and maintain product categories.

### The "Data Utility" Model

The National Retail Data Monitor is a *data utility* for the collection, analysis, and distribution of data on sales of OTC health care products and provision of the data to health departments. This approach reduces the resources required for health departments to monitor sales of OTC health care products. Without a data utility, each health department would have to negotiate data-sharing agreements with many retailers, work with the retailers to understand the data, build systems to collect and analyze the data, and maintain UPC master medication lists as new product codes are assigned. Few health departments have such resources, and retailers have already expressed resistance to this approach. A national data-utility approach also is efficient because most large urban population centers cross jurisdictional boundaries of health departments. Without a centralized approach, each health department would either collect data only for its own jurisdiction (and thus have an incomplete picture of the health of the region) or redundantly collect data for overlapping nearby jurisdictions.

### Data Agreements with Retailers

Working with Information Resources, Inc. (IRI) and ACNielsen, the major syndicated data analysts for U.S. retail industries, the authors identified the significant market share leaders for OTC health care products in the country. The analysis showed that the top five national retailers account for approximately 48% of sales of OTC health care products nationwide, the top 10 retailers for 65%, and the top 20 retailers for 76% of sales. Although the effect of incomplete monitoring of OTC sales on sensitivity of outbreak detection is not fully understood, 65% or 76% sampling far exceeds the sampling efficiency of most public health surveillance schemes.[6,27,28] The process of determining which retailers would be ideally suited for participation involved merging and analyzing a significant amount of industry knowledge and market share data. Figure 1 shows the market coverage provided by the top four national drugstore chains for the 20 most populous U.S. cities. In Washington DC, for example, two retailers account for approximately 75% of all sales of OTC health care products. The chart also identifies the market share of the largest nonnational retailer in the region. In New York City, a large local retailer accounts for an estimated 25% of market share. That retailer is already providing OTC data for surveillance to the New York City Department of Health, and the additional data provided by only two national chains could bring retail data monitoring for the New York City area to over 80%.

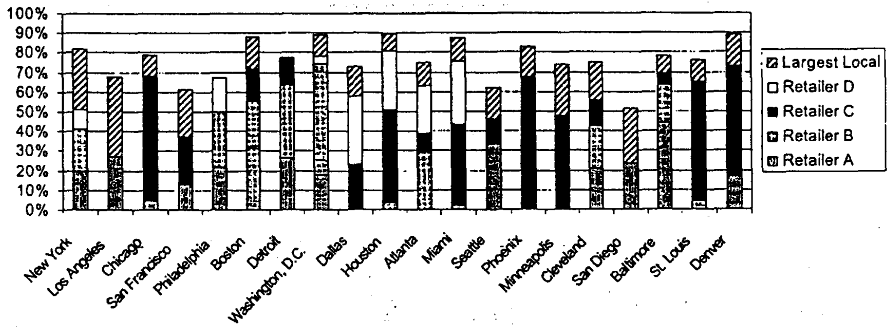


**Figure 1.** Market share of over-the-counter (OTC) health care products of the four largest national retailers for the 20 most populated metropolitan areas in the United States. Also shown is the market share of the largest local retailer for each metropolitan area. The combination of the four national retailers and the largest local retailer provides 50% to 90% market share coverage for all cities. The four retailers do not necessarily correspond to retailers participating in this project. Source: Estimated from industry statistics using ACNielsen Scantrack, IRI Infoscan, and Racher Press Chain Drug Research.

Once the authors identified the retailers with sufficiently large market share to make a meaningful surveillance impact, they asked the retailers to provide their data. Discussions with the retailers were facilitated by introductions made by the IRI and ACNielsen corporations and by letters provided by the Pennsylvania Department of Health and by the Director of the U.S. Centers for Disease Control and Prevention (CDC), Dr. Julie L. Gerberding. Currently, four retailers are providing OTC sales data. For confidentiality reasons, the authors cannot disclose their names, but they comprise approximately 10,000 individual pharmacies representing 23% market share nationally (Fig. 2). Additionally, three other national retailers have agreed to provide data for their 8,000 pharmacies, which will bring the total to approximately 18,000 pharmacies representing approximately 33% of total U.S. OTC health care product sales.

The retailers and the University of Pittsburgh have executed data sharing agreements that permit the University to redistribute the data to those Departments of Health (DOHs) that are participating in the project as well as to the CDC. The agreements stipulate that the data must be aggregated with similar other data by zip code. Each participating DOH is permitted to access only those portions of the data that are relevant to the jurisdiction of that DOH. The DOHs may also review aggregated data using the RODS monitoring system interfaces (described below and in another article in this issue of *JAMIA*). Several retailers expressed concerns that their data not be shared with competitors. These concerns were satisfied by a clause in the data-sharing agreement that prohibits the aggregated data from being shared back to any retailer. Finally, the companies make no warranties of any kind. Retail data that are being collected by the National Retail Data System are not governed by Health Insurance Portability and Accountability Act (HIPAA) regulations because they are not Personal Health Information —the data describe the quantity of product that the stores are selling per day, not what any individual is buying.

## Data Feeds

The requirements for the data feeds from the retailers were minimal time latency and data content with a level of

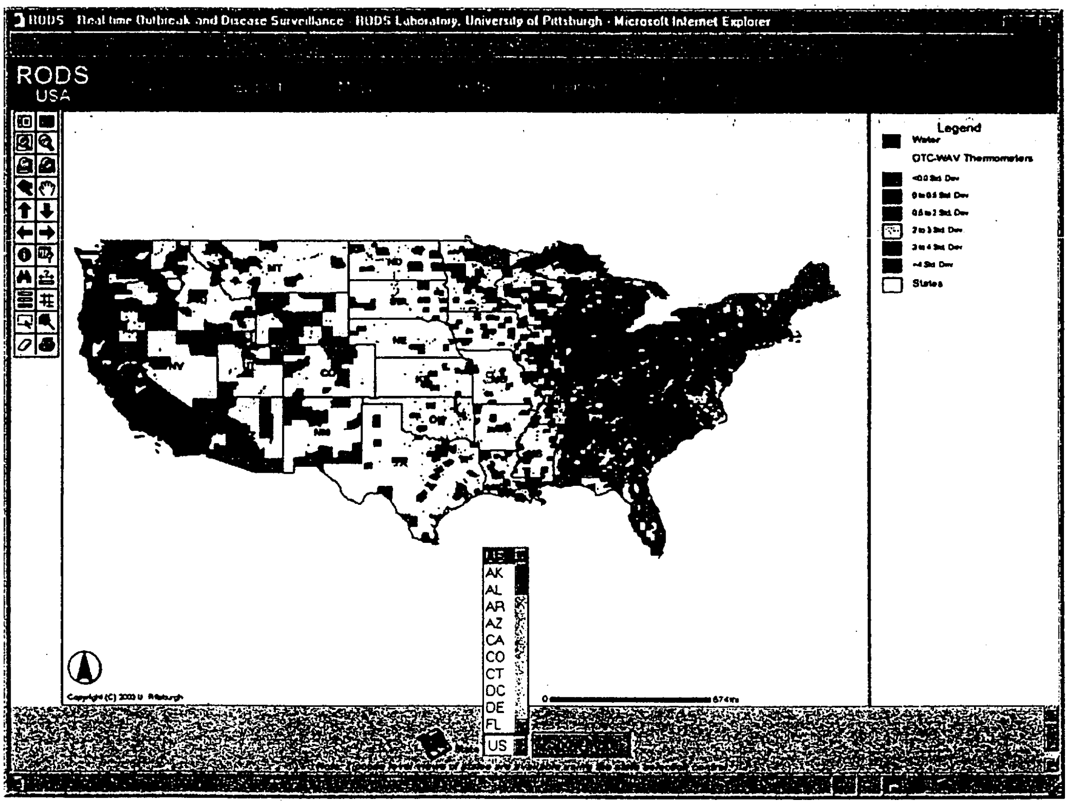


**Figure 2.** National Retail Data Monitor: Sales of thermometers by county on May 17, 2003. This map summarizes data received from approximately 10,000 stores belonging to four national retail chains. The color *green* (the lighter shade) indicates that sales of thermometers on that day were within 0.5 standard deviations of expected, and the color *blue* (the darker shade) indicates sales were between 0.5 and 2.0 standard deviations.

granularity of sales by UPC number, by day, and by store. Due to varying capabilities and preferences, different retailers are at different stages of evolution toward the ideal goal of UPC level of granularity, and this heterogeneity is reflected in the user interface shown in Figure 4. The initial two retailers currently transmit daily counts of OTC sales in five product categories: pediatric electrolytes, cough-and-cold products, thermometers, stomach remedies, and antifever medications. These retailers have agreed to create UPC-level feeds. The third and fourth retailers are transmitting a finer-grained set of categories that contain pediatric-adult distinctions. These retailers have agreed also to create UPC-level feeds. The 18 categories can be aggregated for analytic purposes into the original five categories and are so aggregated in the 'Day Old' categories shown in Figure 4. The fifth, sixth, and seventh retailers—who have agreed to participate but have not yet built data interfaces—will build UPC-level feeds.

Every day, three retailers transfer the previous day's sales data by secure file transfer protocol over the Internet by 3 PM EST. The fourth retailer transmits data approximately every two hours around the clock. The combined data volume currently approximates 14 megabytes (for four retail chains) per day in comma-separated format, although this number will increase when the feeds are converted to UPC-level granularity.

## Data Storage and Security

The data are stored in a secure, firewall-protected facility that has been described previously.[4] The facility meets high availability and capacity requirements. The design assumes that, in the event of a bioterrorism-related outbreak anywhere in the country, thousands of users at health departments will query the site repeatedly and continuously. The technical approach therefore involves (1) fault tolerant network and server configurations; (2) utilization of hardware that supports future mirrored operation at a second site; (3) server clustering that facilitates fault tolerance and load balancing; and (4) early creation of a second mirrored site to ensure against loss through fire or other building event.

## Mapping Universal Product Codes into Product Categories

Many of the distinctions represented by UPC codes have little importance in public health surveillance. It is unlikely that the sale of a 4-ounce bottle of grape-flavored cough syrup is a better indicator of cough than the 8-ounce bottle of cherry-flavored syrup. Moreover, sales of individual products have high variability due to marketing factors such as coupons, discounts, and shelf placement in the store. Therefore, the authors developed a method to aggregate products into analytic classes (*product categories*) for detection. Currently, two different product categorizations are used because there currently is a transition from reporting the original five categories (mentioned earlier) to a new set of 18 more finely grained categories (Table 1). The authors (as clinicians) constructed the 18 categories by reviewing all OTC health care products available on the market across the United States (as defined by the ACNielsen current listing of products). The first step was to eliminate products of no apparent value for outbreak detection such as *sports cream analgesics*. There then remained 7,554 unique OTC health care products (each represented by a unique UPC). Next, each remaining product was assigned to exactly one of the 18 categories by using information that was either explicitly available as a coded characteristic of the product or derivable by simple lexical processing methods from the information provided by ACNielsen. If the latter automated assignment did not work, manual assignment was done based on the same information sources. The assignment process guaranteed that 18 categories were exhaustive and mutually exclusive (i.e., all 7,554 health care products of interest are included, but each only maps to one category). As a result, combinations of the 18 categories can be merged into broader categories (e.g., counts of pediatric and adult cough liquids can be accurately merged into counts of *cough liquids* by simple addition).

*Table 1* ▪ OTC Health Care Product Analytic Categories

| Category | UPC Count |
|---|---|
| Cold relief, adult, liquid | 709 |
| Cold relief, adult, tablet | 2,467 |
| Cold relief, pediatric, liquid | 323 |
| Cold relief, pediatric, tablet | 74 |
| Cough syrup, adult, liquid | 592 |
| Cough syrup, adult, tablet | 32 |
| Cough syrup, pediatric, liquid | 24 |
| Nasal product internal | 371 |
| Throat lozenges | 364 |
| Antipyretic, pediatric | 274 |
| Antipyretic, adult | 1,340 |
| Bronchial remedies | 43 |
| Chest rubs | 78 |
| Diarrhea remedies | 165 |
| Electrolytes, pediatric | 75 |
| Hydrocortisones | 185 |
| Thermometer, pediatric | 125 |
| Thermometer, adult | 313 |
| TOTAL | 7,554 |

## Maintaining UPC-to-category Mappings

Product changes ("new improved formula") result in the assignment of new UPC codes to products, creating a need to update UPC-to-category assignments. Store brands are a particularly difficult instance of this problem, because the retailer, as the manufacturer of the product, has no real need to share an "internal" UPC code outside of the corporation. Because of these considerations, updating UPC-to-category mappings is an important requirement. As a general rule, the majority of UPC code changes in the OTC drug industry occur in August and September when the industry launches new initiatives and releases new products each year. Similar to the technology industry hosting the annual COMDEX trade fair, the drug industry hosts annual food and drug buying "shows" in which the majority of the upcoming merchandising year's buying decisions are made and contracts executed. Fortunately, ACNielsen and IRI monitor UPC-code changes as part of their normal business, and the authors plan to incorporate a similar process for the purpose of ongoing maintenance. The current approach for maintaining UPC data integrity over time compares the existing UPC map with a new universal map from an industry partner on at least a monthly basis to identify new codes. For the future, several experts who represent top manufacturers, retailers, syndicated data analysts, and industry standards governance organizations—including the Uniform Code Council, which manages the distribution of UPCs and all electronic commerce standards in the industry—have agreed to make their expertise available to design a process model that will work throughout the industry.

### Integrating Real-time and Day-old Data

The existence of two different time latencies in the data feeds (three retailers send data at 3 PM about the previous day's sales, and one retailer sends data every two hours) presented a design challenge. Either the real-time data could be held in abeyance until the following day (a lowest common denominator approach), or the real-time data could be handled as a separate data source, facilitating much earlier analysis, with the potential benefit of earlier detection. Because detecting cohort exposures as early as possible is important, the authors chose to treat the real-time data as a separate data stream. Analyzing the data in this manner asks the question, *Can the monitor detect anything unusual about* ***today****, using only data from the real-time stores?* The authors have not yet implemented detection algorithms on the real-time data stream, although the approach will be identical to how the day-old data are analyzed. Currently, detection algorithms operate on the day-old data stream, which includes data from all four retailers. This analysis asks the question, *Can the monitor detect anything unusual about* ***yesterday****, using data from all the stores?*

### User Interfaces and Routing Data to Health Departments

Health departments can access data collected by the National Retail Data Monitor in two ways: (1) through secure Web interfaces that provide temporal and geographic plotting capabilities and (2) through secure raw data feeds that end-user sites can analyze using their own surveillance software or analytic packages. Most health departments use the Web interfaces. Any health department may obtain user accounts for its staff by contacting <nrdmaccounts@cbmi.pitt.edu>. A staff member will be asked to execute a simple data use agreement that limits the use of the data to public health surveillance purposes. Five entities receive raw data on a daily basis at 5 PM EST. They are New York State, New York City, New Jersey, the ESSENCE/JHAP project in the National Capital Area (comprising Washington, DC, Virginia, and Maryland), and the CDC.

Currently, users are instructed to log into the Web interface once a day at 5 PM local time when the data and analyses from the previous day's sales of OTC health care products become available.

#### *Maps*

Users can visualize sales of OTC health care products on maps to detect spatial patterns such as clusters of zip codes with increased sales or linear clusters of zip codes with increased sales. Figure 3 is an example of a map generated by the National Retail Data Monitor. There are currently five such maps (one per product category) per geographic region per day for review. When the monitor is converted to include 18 categories and augmented to map analyses of real-time data, there potentially will be 36 maps to review per day. This number is large. Although some of these product categories may not be of interest to public health and can be eliminated, a more general solution that the authors plan to implement in the near future will be to screen the maps automatically with spatial scan statistics to identify those with anomalies suggesting a need for human review.[29–31]

The maps represent a novel approach to presenting surveillance data. They plot for each zip code—using the colors green, blue, yellow, orange, and red to indicate increasing levels of concern—how "unusual" sales were for the day in question relative to historical patterns of sales for that zip code. In particular, the colors represent the number of standard deviations by which the observed sales of a product category in a zip code deviate from the expected counts. In presenting the data in this fashion, the map serves as a device to focus the user's attention on the degree(s) of anomaly. A user can quickly spot whether the map is predominantly green with a scattering of blue zip codes as would be expected, or whether there are confluent or linear patterns of blue, yellow, orange, or red indicating "unusual" sales activities. The map monitor computes the number of standard deviations relative to a residual signal that has zero mean and constant variation after removal of weekly and longer trends in the data by wavelet transformation. This procedure is intended to produce a "normalized" map that is very sensitive to sudden increases in product counts as would be the case in a medium- to large-scale contamination of the air, food, or water. Alternative transformations of the data are possible using different signal processing approaches focused on detecting more gradual increases. The authors discuss below the reasons that they do not attempt to plot population- and sampling-adjusted mapping.

#### *Normalization Issues*

Because populations and market share coverage for sales of OTC health care products differ between zip codes, plotting raw sales counts is uninformative. The raw sales data for a zip code in midtown Manhattan in the absence of any epidemic might vastly exceed sales in a rural zip code in New Jersey even if every individual in the New Jersey town were to be ill and make purchases.

Epidemiologists are accustomed to adjusting for sampling rate and population differences by plotting the incidence of disease per 100,000 population. Transformations of sales data that approximate this metric would have the desired property of familiarity. If one assumes that the purchase of an OTC health care product is a reasonable indicator of disease (on a one unit to one case basis), one could in theory estimate disease incidence by the following equation that normalizes the raw counts by both market share and population:

$$\textit{incidence} = (\textit{raw daily sales in zip code}/\textit{market share})/\textit{population in zip code}$$

Thus, if the market share being monitored were only 50% in one zip code, after normalizing by market share, the adjusted counts would be double the observed counts from that zip code, making the counts comparable to those from a zip code with 100% market share. After further adjustment by population, the resulting "incidence" would have the desirable property of being comparable across zip codes with different market share and vastly different populations.

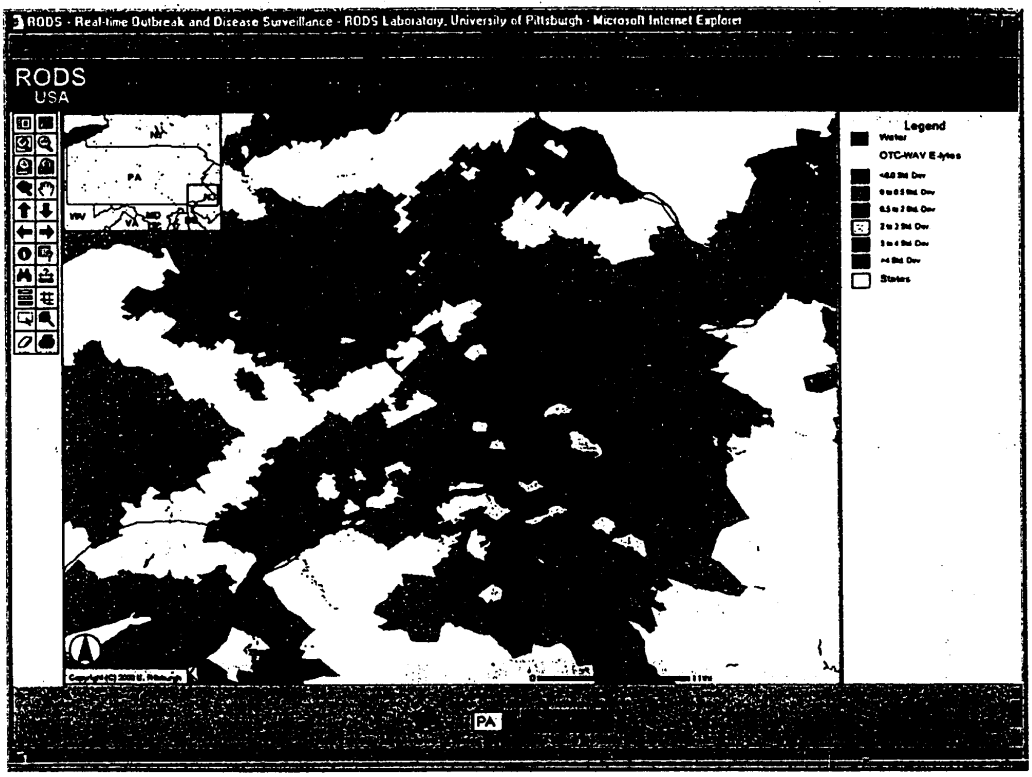


**Figure 3.** National Retail Data Monitor: Sales of pediatric electrolytes in Philadelphia on May 17, 2003. The color coding indicates the level of sales of pediatric electrolytes for each zip code, relative to historical trends for that zip code. Most of the zip codes are colored *green* (*the lighter shade*) indicating that sales are within 0.5 standard deviations of expected. A few zip codes are colored *blue* (*slightly darker shade*), indicating sales are 0.5 to 2.0 standard deviations higher than expected. There are no clusterings of blue zip codes (or yellow, orange, or red areas) that would be indicative of more anomalous sales activity perhaps warranting investigation.

However, there are practical problems that make normalization a nontrivial task. The authors cannot yet obtain good estimates of *market share* at the zip code or even county level. Highly reliable estimates of market share are available at the regional (multicounty) level, but that level of analysis sacrifices spatial granularity. There is a more fundamental problem—people living in a particular zip code may make most of their purchases at stores in a different zip code. For example, commuters to New York City may buy OTC health care products near their place of work, or patrons of a store located on the border of a zip code area may predominantly reside in an adjacent zip code region. For these reasons, the authors took the more general approach to spatial analysis described in the previous section.

### *Time-series Data (Epi Curves)*

Figure 4 shows the epidemic curve plotting capability of the National Retail Data Monitor. The user can review sales of OTC health care products for any product category, region, or time interval. The user can plot raw counts or counts that are normalized (divided) by the total number of OTC health care products sold on that date in the region in question. Normalization in theory is desirable because sales are influenced by nondisease factors such as store hours and bad weather (e.g., blizzards). Such factors potentially could be adjusted for through measures of overall store traffic (as indicated by unique cash register checkouts). The monitor does not yet obtain such measures from all retailers so that the monitor can only normalize by measures of overall sales activities (available as total sales of OTC health care products). A problem with this approach is that sales of OTC health care products are dominated by the cough-and-cold category, so normalization of the cough-and-cold signal itself by total sales will tend to remove any real spikes in cough-and-cold sales. For this reason, the authors recommend that users look at raw data in the current interface.

Store and newspaper product promotions are another potential confounder, although retailers know that it is very difficult to promote increased consumption in the absence of disease for many products. Promotions may affect which specific products consumers purchase within a category but

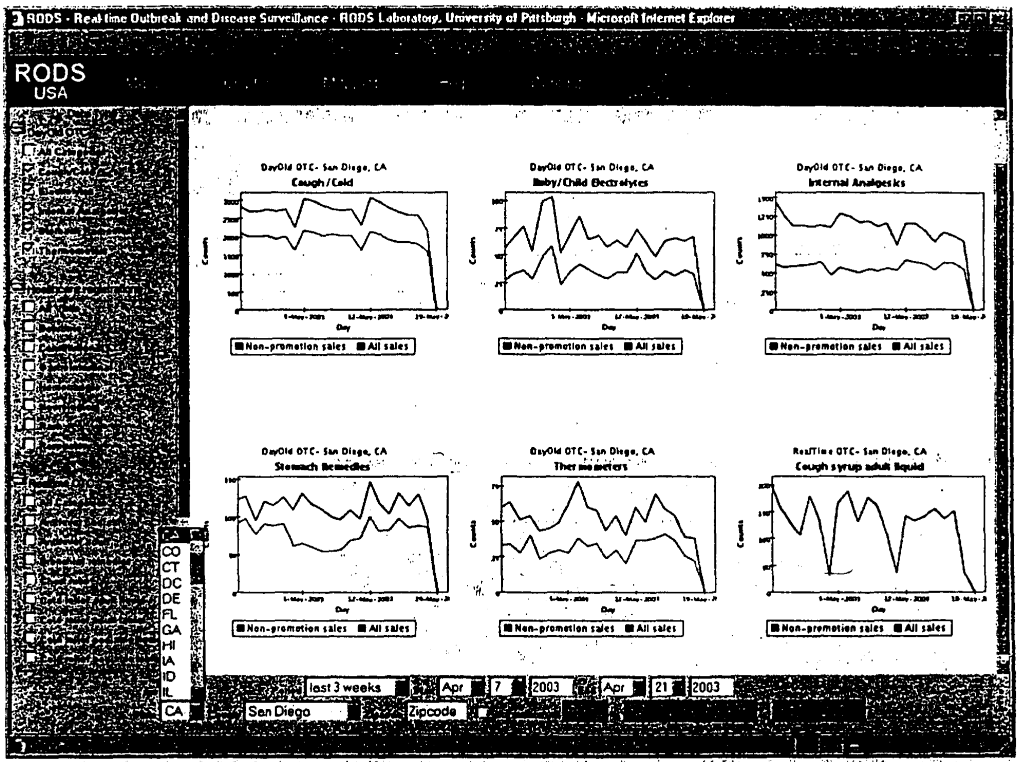

RODS - Real-time Outbreak and Disease Surveillance - RODS Laboratory, University of Pittsburgh - Microsoft Internet Explorer
RODS
USA
DayOld OTC- San Diego, CA
Cough/Cold
DayOld OTC- San Diego, CA
Baby/Child Electrolytes
DayOld OTC- San Diego, CA
Internal Analgesics
DayOld OTC- San Diego, CA
Stomach Remedies
DayOld OTC- San Diego, CA
Thermometers
RealTime OTC- San Diego, CA
Cough syrup adult liquid
Counts
Day
■ Non-promotion sales ■ All sales
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
CA
last 3 weeks
Apr
7
2003
Apr
21
2003
San Diego
Zipcode

**Figure 4.** National Retail Data Monitor: Sales of six product categories in San Diego for a three-week period. This screen shows daily sales of five *DayOld OTC* product categories and the *Cough syrup adult liquid Realtime OTC* category (*bottom right of screen*). The upper line in each graph represents total sales and the lower, nonpromotional. A dip in Sunday sales of cough and cold products is visible. Users can view an arbitrary number of such graphs by checking the desired graphs on the list on the left of the screen and then clicking "Plot It." At the bottom of the screen are controls for selecting the geographic region, time interval, and normalization (by total OTC sales). Controls also allow download of 30 days of data to a file for off-line analysis. The "Get Cases" function is not available for OTC sales data.

do not affect category-level sales. Nevertheless, the data feeds from retailers distinguish between sales of promoted and nonpromoted items (i.e., there are two records for every category—that day's sales of promoted products and that day's sales of nonpromoted product). To expose the potential effects of promotion on the time series, the user interfaces plot both nonpromoted sales and total sales (the lower and upper lines, respectively, in the graphs in Figure 4).

## Status Report

The National Retail Data Monitor has been in continuous operation since December 2002. The authors consider the project to be in its developmental "build" phase due to ongoing work with the retail industry to achieve 70% data coverage, reduce time latency, and move to UPC-level data feeds. As of May 15, 2003, the data coverage of the system is 23% of total national sales of OTC health care products. The time latency is one day or less. The project has created 119 user accounts for health department employees in 17 states, Washington DC, and at the CDC. Five entities receive raw data feeds from the system including New York City, JHAP/ Essence in the National Capital Area, New York State, New Jersey, and the CDC.

The goal for system usage is that the product category geographic maps for each jurisdiction be reviewed every day when they become available at 5 PM local time. This map-based analysis is a relatively recent feature and there has not been sufficient time for education about the schedule of availability and its proper use. Recent usage statistics indicate that on weekdays, at least 15 unique users log in per day, dropping to three on weekends. The authors have not analyzed this usage data sufficiently to understand how many jurisdictions are monitoring the data routinely. Ideally, differential rates for weekend and weekday usage should not exist in an operational surveillance system. The data suggest that, possibly related to temporal staffing patterns, some users may perceive the system as something to consult in the event of an epidemic or a heightened level of alert but not as an early warning system (its intended use). The long-term status of the system is under discussion with a coalition of states that are considering creating shared surveillance resources. During the ongoing "build" phase, the National Retail Data Monitor will continue to be supported by grant funding from the Commonwealth of Pennsylvania. The authors have actively sought additional support from foundations, the Department of Homeland Security, Department of Health and Human Services, and CDC.

## Discussion

When completed, the National Retail Data Monitor will make available to health departments, in near to real time, UPC-level product sales data in both raw and analyzed form representing at least 70% of sales in their jurisdictions. These data may be useful for public health surveillance for bioterrorism, infectious diseases, and chronic diseases. The data will also be useful for prospective validation studies of new detection methods. The National Retail Data Monitor represents a model for developing a "data utility" that can serve public health surveillance. As such, the authors' experiences might facilitate future development of public health surveillance tools using data from nurse call centers, health maintenance organizations, national laboratory companies, and poison call centers.

The authors found that a key element for success included the deep understanding of the industry provided to them by an industry expert. This knowledge was invaluable in crafting all aspects of the project, from the "80–20 sufficing" approach, to obtaining the data, to designating product categories, and to maintenance issues. The authors also found that presenting the scientific case for the value of the data was important. Equally key was a personal invitation, sent to the CEO of relevant corporations, for participation (sharing of otherwise proprietary data), authored by a highly respected government or public health official. Another success factor involved the development of an interdisciplinary team with expertise in medical informatics, computer science, law, and engineering.

Many of the problems encountered in analysis and presentation of surveillance data are not unique to retail data. For example, the problem of merging similar data arriving from multiple sources with different time latencies will be common in new surveillance approaches (e.g., hospital registration data may be available in real time from some sources and in batch mode with a one-day delay from other sources). The issues of normalization for spatial and temporal analysis are generic, and we have already encountered them with analysis of hospital chief complaint data in our own work and anticipate that they will be found in analysis of call data to nurse call lines. The method of spatial analysis involving plotting standard deviations for each spatial cell, based on that cell's historical expectation, may be widely applicable in other domains.

Other lessons learned include that data sharing agreements should allow redistribution of data to any public health authority and permit data to be used in research. Data sources that are amenable to a "national" approach should be formed into data utilities—services independent of any particular user interface—and should be industry based. Health departments differ in their needs. Some agencies prefer to receive raw data because they already have surveillance "front ends" or are more comfortable using "off-line" analytic packages. Although there is very significant value in creating analytic applications for users, delivering the data through only a monolithic, vertical application would not meet the needs of a significant subset of end-users (health departments).

The project's immediate future plans are to achieve the target of 70% data coverage and to reduce time latencies toward "real time." The project intends to add a rapid spatial scan algorithm to automate more fully the review of maps. The project will develop an evaluation strategy to document prospectively the system's ability to detect naturally occurring outbreaks, should they occur.

Longer-term project plans include the expansion of monitoring to the level of selected prescription medications. As in the case of retail products, a standard coding system (NDC) that is used in industry data systems provides the basis of feasibility. The national retailers' store prescription data are stored separately from retail sales data. Prescription data are subject to HIPAA controls, but aggregation of the data by zip code, similar to how sales of OTC health care products are aggregated, should satisfy HIPAA regulations.

Future plans also include extension of the project to international scope. Just as the Uniform Code Council administers and manages UPCs in the United States and Canada, EAN International provides the same service outside of North America by using European Article Numbering (EAN) codes. EAN is a UPC-compatible system, which began in the 1970s and eventually merged into the current EAN·UCC System, used as a standard by 45 countries including Japan, and most of the European Union for coding pharmaceutical products.

## Conclusions

The National Retail Data Monitor is a general model for a class of surveillance approaches that leverage existing commercial data collections. The promise of such systems derives from the inherent early availability of their data (reflecting early illness behaviors) and the extreme efficiency with which the data can be obtained, relative to more traditional surveillance data methods. Only with real-world experience in detecting various types of outbreaks will the true utility of the National Retail Data Monitor become known.

*References* ■

1. Wagner M, Tsui F-C, Espino J, et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract. 2001;6(6):50–8.
2. Kaufmann A, Meltzer M, Schmid G. The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? Emerg Infect Dis. 1997;3(2):83–94.
3. Lober WB, Karras BT, Wagner MM, et al. Roundtable on bioterrorism detection: information system-based surveillance. J Am Med Inform Assoc. 2002;9:105–15.
4. Tsui F-C, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: a real-time public health surveillance system. J Am Med Inform Assoc. 2003;10: 399–408.
5. Talan DA, Moran GJ, Mower WR, et al. EMERGEncy ID NET: an emergency department-based emerging infections sentinel network. The EMERGEncy ID NET Study Group. Ann Emerg Med. 1998;32:703–11.

6. Effler P, Ching-Lee M, Bogard A, et al. Statewide system of electronic notifiable disease reporting from clinical laboratories: comparing automated reporting with conventional methods. JAMA. 1999;282:1845–50.
7. Lewis M, Pavlin J, Mansfield J, et al. Disease outbreak detection system using syndromic data in the greater Washington DC area. Am J Prev Med. 2002;23:180.
8. Lazarus R, Kleinman K, Dashavsky I, et al. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerg Infect Dis. 2002;8:753–60.
9. Lazarus R, Kleinman KP, Dashavsky I, et al. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health. 2001;1(1):9.
10. Zelicoff A, Brillman J, Forslund DW, et al. The Rapid Syndrome Validation Project (RSVP). Proc AMIA Symp. 2001:771–5.
11. Gesteland PH, Wagner MM, Chapman WW, et al. Rapid deployment of an electronic disease surveillance system in the State of Utah for the 2002 Olympic Winter games. Proc AMIA Symp. 2002:285–9.
12. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health clinical medicine. J Public Health Manag Pract. 2001;7(6):43–50.
13. Labrie J. Self-care in the new millennium: American attitudes towards maintaining personal health. Consumer Healthcare Products Association, 2001, p 76. <http://www.chpa-info.org/pdfs/CHPA%20Final%20Report%20revised%20(03-20)_.pdf>. Accessed July 12, 2003.
14. McIsaac WJ, Levine N, Goel V. Visits by adults to family physicians for the common cold. J Fam Pract. 1998;47:366–9.
15. Hogan WR, Tsui F-C, Ivanov O, et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. J Am Med Inform Assoc. 2003;10 (in press).
16. Magruder S, Florio E. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of public health. Johns Hopkins University Applied Physics Laboratory Technical Digest. 2003;24(4) (in press).
17. Welliver RC, Cherry JD, Boyer KM, et al. Sales of nonprescription cold remedies: a unique method of influenza surveillance. Pediatr Res. 1979;13:1015–7.
18. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc Natl Acad Sci U S A. 2002;99:5237–40.
19. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Med Inform Decis Mak. 2003;3(1):2.
20. Broome C, Pinner R, Sosin D, Treadwell T. On the threshold. Am J Prev Med. 2002;23(1):229.
21. Halperin W, Baker ELJ (eds). Public Health Surveillance. New York: Van Nostrand Reinhold, 1992.
22. Teutsch S, Churchill R. Principles and Practice of Public Health Surveillance (ed 2). Oxford: Oxford University Press, 2000.
23. Yasnoff WA, Overhage JM, Humphreys BL, et al. A national agenda for public health informatics. J Public Health Manag Pract. 2001;7(6):1–21.
24. Goldstein A. Strategic tracking of sniffles: scientists on alert for terrorism monitor area health factors. Washington Post. March 28, 2003: A10. <http://www.washingtonpost.com/wp-dyn/articles/A39629-2003Mar27.html>. Accessed July 12, 2003.
25. Perez-Peña R. An early warning system for diseases in New York. New York Times. April 4, 2003. <http://www.nytimes.com/2003/04/04/nyregion/04WARN.html?ex=1053489600&en=a50eec53d50d3da6&ei=5070>. Accessed May 18, 2003.
26. Wein LM, Craft DL, Kaplan EH. Emergency response to an anthrax attack. Proc Natl Acad Sci U S A. 2003;100:4346–51.
27. Panackal AA, M'ikanatha NM, Tsui F-C, et al. Automatic electronic laboratory-based reporting of notifiable infectious diseases. Emerg Infect Dis. 2001;8:685–91.
28. Ewert DP, Westman S, Frederick PD, Waterman SH. Measles reporting completeness during a community-wide epidemic in inner-city Los Angeles. Public Health Rep. 1995;110:161–5.
29. Kulldorff M, Athas WF, Feurer EJ, et al. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health. 1998;88:1377–80.
30. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med. 1995;14:799–810.
31. Glaz J, Balakrishnan N. Scan statistics and applications. Statistics for industry and technology. Boston: Birkhèauser, 1999, pp xxi, 324.

The Practice of Informatics

*Application of Information Technology* ■

# Technical Description of RODS: A Real-time Public Health Surveillance System

Fu-Chiang Tsui, PhD, Jeremy U. Espino, MD, Virginia M. Dato, MD, MPH, Per H. Gesteland, MD, MS, Judith Hutman, Michael M. Wagner, MD, PhD

**Abstract** This report describes the design and implementation of the Real-time Outbreak and Disease Surveillance (RODS) system, a computer-based public health surveillance system for early detection of disease outbreaks. Hospitals send RODS data from clinical encounters over virtual private networks and leased lines using the Health Level 7 (HL7) message protocol. The data are sent in real time. RODS automatically classifies the registration chief complaint from the visit into one of seven syndrome categories using Bayesian classifiers. It stores the data in a relational database, aggregates the data for analysis using data warehousing techniques, applies univariate and multivariate statistical detection algorithms to the data, and alerts users of when the algorithms identify anomalous patterns in the syndrome counts. RODS also has a Web-based user interface that supports temporal and spatial analyses. RODS processes sales of over-the-counter health care products in a similar manner but receives such data in batch mode on a daily basis. RODS was used during the 2002 Winter Olympics and currently operates in two states—Pennsylvania and Utah. It has been and continues to be a resource for implementing, evaluating, and applying new methods of public health surveillance.

■ J Am Med Inform Assoc. 2003;10:399–408. DOI 10.1197/jamia.M1345.

Covert, large-scale attacks using biological agents such as anthrax, plague, tularemia, and smallpox can lead to massive casualties unless quarantine, vaccination, and/or antibiotic treatments are instituted promptly.[1] History highlights the need for timely detection of these threats. In 1979, there was an accidental release of anthrax from a bioweapons plant in Sverdlovsk, Russia. Before the anthrax outbreak was recognized, at least six patients exhibited influenzalike symptoms and were dismissed by their physicians as not having any serious illness. Twenty-one individuals had already died by the time laboratories confirmed the presence of *Bacillus anthracis*.[2]

Unfortunately, conventional public health disease surveillance—which relies on physician and laboratory reporting and manual analysis of surveillance data—is ill equipped for timely detection of such threats.[3] The reportable disease system relies on health care professionals to recognize, diagnose, and report cases and suspected outbreaks to public health officials[4,5]; however, it is unlikely that without an event or alert to raise his or her index of suspicion, a physician will attribute the early symptoms and signs of disease in a bioattack victim appropriately and report the case.[6] A key limitation of the current system is that the lone physician is blind to the cases his or her colleagues in a nearby hospital are seeing—knowledge that might lead the physician to consider uncommon diseases more strongly in his or her diagnostic reasoning. Mandatory laboratory reporting[4] is also ill-equipped for early detection, because it takes time before tests are ordered and specimens are obtained, transported, processed, and resulted.

Sufficiently early detection of a biological attack may be accomplished through surveillance schemes that can detect infected individuals earlier in the disease process. For completeness, we note that biosensors are being developed (and deployed) that detect organisms in the air and that this type of detection, if feasible, occurs fundamentally much earlier, because the delay introduced by the incubation period

Affiliations of the authors: University of Pittsburgh, Pittsburgh, Pennsylvania (F-CT, JUE, VMD (formerly), MMW); University of Utah, Salt Lake City, Utah (PHG); Intermountain Health Care, Salt Lake City, Utah (PHG). Dr. Dato is currently employed by the Pennsylvania Department of Health.

This work was supported in part by grants GO8 LM06625-01 and T15 LM/DE07059 from the National Library of Medicine; contract no. 290-00-0009 from the Agency for Healthcare Research and Quality; and by Cooperative Agreement Number U90/CCU318753-02 from the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsiblity of the authors and do not necessarily represent the official views of CDC. Hoah-Der Su, Zhen Liu, and Xiaoming Zeng wrote components of RODS. The authors are indebted to Kimberlee Barnhart, Mary Cleatus Szczepaniak, participating hospitals, interface and network engineers, ESRI Inc., Bill Davenhall, and Lori Shienvold.

Correspondence and reprints: Fu-Chiang Tsui, PhD, The RODS Laboratory, Center for Biomedical Informatics, Suite 8084 Forbes Tower, 200 Lothrop Street, Pittsburgh, PA 15261; e-mail: <tsui@cbmi.pitt.edu>.

Received for publication: 02/19/03; accepted for publication: 05/13/03.

of the. disease is eliminated from the surveillance system.[7] However, such approaches face unsolved technical problems in the analysis of contaminated specimens (the norm in air sampling). Biosensors also need to be in the right place—on every person's lapel or every street corner and hallway—to provide complete surveillance coverage.

Surveillance methods that can detect disease at an earlier stage are an important research direction for public health surveillance. These methods are generally referred to as *syndromic surveillance* because they have the goal of recognition of outbreaks based on the symptoms and signs of infection and even its effects on human behavior prior to first contact with the health care system.[8] Because the data used by syndromic surveillance systems cannot be used to establish a specific diagnosis in any particular *individual*, syndromic surveillance systems must be designed to detect signature *patterns of disease in a population* to achieve sufficient specificity. For example, it would be absurd to use only the symptom of fever to attempt to establish a working diagnosis of inhalational anthrax in an individual, but it would be very reasonable to establish a working diagnosis of anthrax release in a community if we were to observe a pattern of 1,000 individuals with fever distributed in a linear streak across an urban region consistent with the prevailing wind direction two days earlier. It would be beyond reasonable and, in fact, imperative to establish a working diagnosis of public health emergency if presented with such information.

One recent example of a form of syndromic surveillance is *drop-in surveillance*—the stationing of public health workers in emergency departments (EDs) and special clinics during high-profile events such as the Super Bowl to capture data on patients presenting with symptoms potentially indicative of bioterrorism. The major disadvantage of this approach is the cost of round-the-clock staffing for manual data collection.

A less expensive approach—and the one taken in the Real-time Outbreak and Disease Surveillance (RODS) system—is detection based on data collected routinely for other purposes. Examples of such data include absenteeism data, sales of over-the-counter (OTC) health care products, and chief complaints from EDs.[9] The expenses of manual data collection are avoided; however, the data obtained typically are noisy approximations of what could be obtained by direct interviewing of the patient (in the case of individual level data). Both approaches may play complementary roles with current methods of public health surveillance[10–12] by assisting the physician and public health official with a continuously updated picture of the "health status" of a population.[13,14]

A focus of our research has been syndromic surveillance from free-text chief complaints routinely collected by triage nurses in EDs and acute care clinics during patient registration. We have deployed this type of surveillance at the 2002 Winter Olympics and in the States of Pennsylvania and Utah. We described a previous version of the RODS system,[12] but the system has undergone considerable subsequent development both architecturally and functionally. This report provides a detailed description of the current version of RODS, an example of a computer-based public health surveillance system that adheres to the National Electronic Disease Surveillance System (NEDSS) specifications of the Centers for Disease Control and Prevention (CDC).[15,16]

## Background

### Public Health Surveillance

The role of public health surveillance is to collect, analyze, and interpret data about biological agents, diseases, risk factors, and other health events and to provide timely dissemination of collected information to decision makers.[17] Conventionally, public health surveillance relies on manual operations and off-line analysis.

### Syndromic Surveillance

Existing syndromic surveillance systems include the CDC's drop-in surveillance systems,[8] Early Notification of Community-based Epidemics (ESSENCE),[10,18] the Lightweight Epidemiology Advanced Detection and Emergency Response System (LEADERS),[19] the Rapid Syndrome Validation Project (RSVP),[20] and the eight systems discussed by Lober et al.[11]

Lober et al. summarized desirable characteristics of syndromic surveillance systems and analyzed the extent to which systems that were in existence in 2001 had those characteristics.[11] A limitation[7] of most systems (e.g., ESSENCE,[10] Children's Hospital in Boston,[11] University of Washington[11]) was batch transfer of data, which may delay detection by as long as the time interval (periodicity) between batch transfers. For example, a surveillance system with daily batch transfer may delay by one day the detection of an outbreak.

Some systems required manual data input (e.g., CDC's drop-in surveillance systems, RSVP,[20] and LEADERS[19]), which is labor-intensive and, in the worst case, requires round-the-clock staffing. Manual data input is not a feasible mid- or long-term solution even if the approach is to add items to existing encounter forms (where the items still may be ignored by busy clinicians).

A third limitation for existing surveillance systems is that the systems may not exploit existing standards or communication protocols like Heath Level 7 (HL7) even when they are available.

The data type most commonly used among surveillance systems is symptoms or diagnoses of patients from ED and/or physician office visits. Other types of data identified in that study include emergency call center and nurse advice lines. Other types of data being used include sales of over-the-counter health care products, prescriptions, telephone call volumes to health care providers and drug stores, and absenteeism. We have conducted studies demonstrating that the free-text chief complaint data that we use correlate with outbreaks.[21,22]

## Design Objectives

The overall design objective for RODS is similar to that of an early warning system for missile defense; namely, to collect whatever data are required to achieve early detection from as wide an area as necessary and to analyze the data in a way that they can be used effectively by decision makers. It is required that this analysis be done in close to real time. This design objective is complex and difficult to operationalize because of the large number of organisms and the even larger number of possible routes of dissemination all requiring potentially different types of data for their detection, different algorithms, and different time urgencies. For this reason, our focus since beginning the project in 1999 has been on the specific problem of detecting a large-scale outbreak due to an outdoor (outside buildings) aerosol release of anthrax. Additional design objectives were adherence to NEDSS standards to ensure future interoperability with other types of public health surveillance systems, scalability, and that the system could not rely on manual data entry, except when it was done in a focused way in response to the system's own analysis of passively collected data.

## Technical Description

This report describes RODS 1.5, which was completely rewritten as a Java 2 Enterprise Edition (J2EE) application since the previous publication describing it. RODS 1.5 is multidata type enabled, which means that any time series data can be incorporated into the databases and user interfaces. The deployed RODS system currently displays and analyzes health care delivery site registrations and separately monitors sales of OTC health care products.

### Overview

RODS uses clinical data that are already being collected by health care providers and systems during the registration process. When a patient arrives at an ED (or an InstaCare in Utah), the registration clerk or triage nurse elicits the patient's reason for visit (i.e., the *chief complaint*), age, gender, home zip code, and other data and enter the data in a registration computer. The registration computer then generates an HL7 ADT (admission, discharge, and transfer) message and transmits it to the health system's HL7 message router (also called an *integration engine*). There usually is only one message router per health system even if there are many hospitals and facilities. These processes are all routine existing business activities and do not need to be created de novo for public health surveillance.

Figure 1 shows the flow of clinical data to and within RODS. The hospital's HL7 message router, upon receipt of an HL7 message from a registration computer, deletes identifiable information from the message and then transmits it to RODS over a secure virtual private network (VPN), or a leased line, or both (during the 2002 Winter Olympics we utilized both types of connections to each facility for fault tolerance). The RODS HL7 listener maintains the connection with the health system's message router and parses the HL7 message as described in more detail below. It then passes the chief complaint portion of the message to a Bayesian text classifier that assigns each free-text chief complaint to one of seven syndromic categories (or to an eighth category, *other*). The database stores the category data, which then are used by applications such as detection algorithms and user interfaces.

Data about sales of OTC health care products are processed separately by the National Retail Data Monitor, which is discussed in detail in another article in this issue of *JAMIA*.[23] The processing was kept separate intentionally because, in the future, the servers for the National Retail Data Monitor may operate in different physical locations than RODS. The RODS user interfaces can and do display sales of OTC health care products as will be discussed, but other user interfaces can be connected to the National Retail Data Monitor as well.

### Data Level

#### *Data Sharing Agreements*

Prior to September 2001, RODS received data only from hospitals associated with the UPMC Health System, and efforts to recruit other hospitals met with resistance. After the terrorist attacks (including anthrax) in the Fall of 2001, other hospitals agreed to participate. Although data in this project are de-identified, certain information such as the number of ED visits by zip code were considered proprietary information by some health systems. Health Insurance Portability and Accountability Act (HIPAA) concerns also were very prominent in the discussions. Data-sharing agreements were

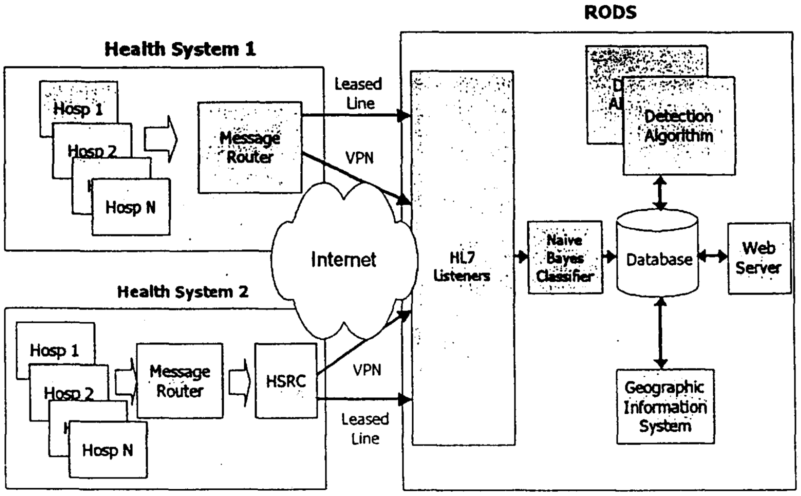


**Figure 1.** Data communication to RODS system from various health systems. (HSRC, health system resident component; VPN, virtual private network).

executed with every participating health system that addressed these concerns. As an additional precaution, all RODS project members meet annually with University of Pittsburgh council to review obligations and are required to sign an agreement every year stating that they understand the terms of the data-sharing agreements and agree to abide by the terms. RODS began as a research project at the University of Pittsburgh in 1999 and has functioned with IRB approvals since that time.

### *Data Types*

Health care facilities send admission, discharge, and transfer (ADT) HL7 messages to RODS for patient visits in EDs and walk-in clinics. A minimal data set is sent, as shown in Figure 2, which qualifies as a HIPAA Limited Data Set.[24] Currently the data elements are age (without date of birth), gender, home zip code, and free-text chief complaint.

### *Data Transmission*

The HL7 listener receives HL7 messages from the message routers located in each health system. The HL7 listener then passes the received HL7 message to the HL7 parser bean, an Enterprise JavaBean (EJB) in the RODS business logic tier. The HL7 parser bean uses regular expressions to parse the fields in an HL7 message. The HL7 parser bean then stores the parsed elements into a database through a managed database connection pool.

Although nearly all health systems utilize the HL7 messaging standard, the location of individual data elements in an HL7 message may differ from health system to health system. For example, some care providers' systems record free-text chief complaint in the DG1 segment instead of the PV2 segment of an HL7 message. To resolve this mapping problem, a configuration file written in eXtensible Markup Language (XML), a standard protocol often used to define hierarchical data elements, defines where each of the data elements can be found in the HL7 message. When an HL7 listener starts up, it reads the hospital-dependent configuration file and passes the configuration information to the parser bean.

We also use this configuration file to define the database table and field in which the HL7 parser bean should store each data element. This approach is useful because it allows the HL7 data to be stored to an external database. We anticipate that health departments with existing NEDSS or other public health surveillance databases may wish to use just this component of RODS for real-time collection of clinical data.

For hospitals that do not have HL7 message routers (two of approximately 60 in our experience to date), RODS accepts ED registration data files through either a secure Web-based data upload interface or a secure file transfer protocol. In general, these types of data transfers are technically trivial and for that reason are used by many groups but do not have the reliability of a HL7 connection (and have very undesirable time latencies).

### *Data Integrity*

RODS checks the integrity of the data in the HL7 messages that it receives. This processing is necessary because hospital data flows may have undesirable characteristics such as duplicates. RODS identifies and deletes duplicates by using a database trigger that creates a composite primary key before inserting the data. RODS also filters out scheduling messages, which are identified by the fact that they have *future* admitted date and time.

RODS monitors all data feeds to ensure continuous connections with health systems. If RODS does not receive data for six hours, it sends an alert to the RODS administrator and the sending health system's administrator. Because the commercial message routers that hospitals use queue up HL7 messages when encountering networking or system problems, data integrity is preserved.

### *Database*

RODS uses an Oracle8i database to store ED registration data. (Oracle, Redwood Shores, CA). To ensure fast response for an online query (e.g., the daily counts of respiratory syndrome in a county for the past six months), we developed a cache table scheme that pre-aggregates counts and refreshes them every 30 minutes.[25]

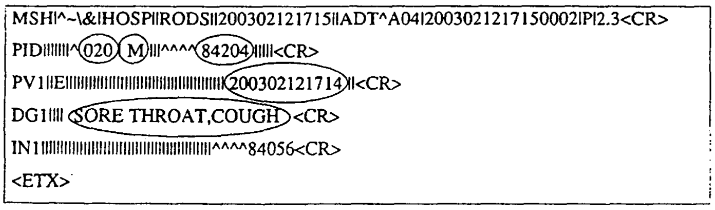
## Network Level

The communications network between RODS and health care systems consists of virtual private networks (VPN) and leased lines. RODS uses multivendor site-to-site Internet Protocol Security (IPSEC) VPNs to receive HL7 messages. During the Winter Olympics, we exclusively used leased lines for the primary connection because of concerns about possible communications interruptions due to Internet traffic related to the games. The leased lines consisted of a redundant pair of 128k fractional T1 lines. After the Olympics, we returned to use of VPNs, and RODS has operated reliably using VPNs in both Utah and Pennsylvania. The leased-line modality is used only to connect the Siemens Medical Systems Data Center with RODS for the transmission of data from nine health systems that are hosted by Siemens.

## System Hardware

For connectivity with the HL7 message routers, we utilize hardware-based routers. The VPN router is a Cisco PIX 501 and the leased-line routers are a pair of Cisco 2600s (Cisco Systems, Inc., San Jose, CA).

All of the RODS processes can be run on a single computer, but in our current implementation—serving Pennsylvania



```
MSH|^~\&|HOSP||RODS||200302121715||ADT^A04|2003021217150002|P|2.3<CR>
PID|||||||^020M|||^^^^84204|||||<CR>
PV1||E||||||||||||||||||||||||||||||||||||||||||200302121714||<CR>
DG1|||| SORE THROAT,COUGH <CR>
IN1||||||||||||||||||||||||||||||||||||||||||||||^^^^84056<CR>
<ETX>
```

**Figure 2.** Sample HL7 admission, discharge, and transfer (ADT) message from an emergency department. The circled fields are age, gender, home zip code, admitted date and time, and free-text chief complaint, respectively.

and Utah as an application service provider—we use five dedicated servers: firewall, database, Web server, a geographic information system (GIS) server, and computation. The processes are written in Java code and can run on most platforms, but here we describe the specific platforms we use to indicate approximate sizing and processing requirements.

The database server is a Sun Microsystems Enterprise 250 configured with two Ultrasparc II 400Mhz processors, 2 gigabytes of RAM, and 36 gigabytes of mirrored hard drive space running an Oracle 8.1.7 (database) on Solaris 8 (Sun Microsystems, Inc., Santa Clara, CA).

The Web server is a Dell Poweredge 1550 configured with two 1Ghz Pentium III processors, 1 gigabyte of RAM, and 36 gigabytes of Redundant Arrays of Inexpensive Disk 5 (RAID-5) storage running Apache 1.3.24 (Web server), and Jboss 3.0 (described below in Fault Tolerance) on Redhat Linux 7.1 (Dell Computer Corporation, Round Rock, TX; Jboss Group, Atlanta, GA; Red Hat, Raleigh, NC).

The GIS server is a Dell Poweredge 350 configured with one 1Ghz Pentium III processor, 512 megabytes of RAM, and 18 gigabytes of storage running ArcIMS 4.0 (ESRI, Inc., Redlands, CA), an Internet-enabled geographic information system on Redhat Linux 7.3.

The computation server is a Penguin Computing server configured with dual Athlon MP 2400s, 1 gigabyte of RAM, and 750 gigabytes of RAID-5 storage running Oracle 9i on Redhat Linux 7.3.

Backup is performed nightly on all machines using a Sun StoreEdge L9 Tape Autoloader attached to the database server and Veritas Netbackup software (Veritas, Mountain View, CA).

## Application Level

We developed RODS applications using the Java 2 Enterprise Edition Software Toolkit (J2EE SDK) from Sun Microsystems for cross-platform Java application development and deployment.[26]

We followed contemporary application programming practices—a multitiered application consisting of a client tier (custom applications such as HL7 listeners and detection algorithms), business logic tier, database tier, and Web tier.

Business logic such as the HL7 parser bean was implemented as Enterprise JavaBeans (EJBs). NEDSS specifies EJB as the standard for application logic. RODS uses Jboss, an open-source J2EE application server, to run all EJBs.[10]

The Web tier comprises the graphical user-interface to RODS and uses Java Server Pages (JSP), Java Servlets, and ArcIMS. The database tier was implemented in Oracle 8i.

### *Natural Language Processing*

RODS uses a naive Bayesian classifier called *Complaint Coder* (CoCo) to classify free-text chief complaints into one of the following syndromic categories: constitutional, respiratory, gastrointestinal, neurological, botulinic, rash, hemorrhagic, and other. CoCo computes the probability of each category, conditioned on each word in a free-text chief complaint and assigns a patient to the category with the highest probability.[27] The probability distributions used by CoCo are learned from a manually created training set. CoCo can be retrained with local data, and it can be trained to detect a different set of syndromes than we currently use. CoCo runs as a local process on the RODS database server. CoCo was developed at the University of Pittsburgh and is available for free download at <http://health.pitt.edu/rods/sw>.

### *Detection Algorithms*

Over the course of the project, RODS has used two detection algorithms. These algorithms have not been formally field tested because the emphasis of the project to date has been on developing the data collection infrastructure more than field testing of algorithms.

The Recursive-Least-Square (RLS) adaptive filter[28] currently runs every four hours, and alerts are sent to public health officials in Utah and Pennsylvania. RLS, a dynamic autoregressive linear model, computes an expected count for each syndrome category for seven counties in Utah and 16 counties in Pennsylvania as well as for the combined counts for each state. We use RLS because it has a minimal reliance on historical data for setting model parameters and a high sensitivity to rapid increases in a time series e.g., a sudden increase in daily counts. RLS triggers an alert when the current actual count exceeds the 95% confidence interval for the predicted count.

During the 2002 Olympics we also used the What's Strange About Recent Events (WSARE 1.0) algorithm.[29] WSARE performs a heuristic search over combinations of temporal and spatial features to detect anomalous densities of cases in space and time. Such features include all aspects of recent patient records, including syndromal categories, age, gender, and geographical information about patients. The criteria used in the past for sending a WSARE 1.0 alert was that there has been an increase in the number of patients with specific characteristics relative to the counts on the same day of the week during recent weeks and the p-value after careful adjustment for multiple testing for the increase was ≤0.05. Version 3.0 of WSARE, which will incorporate a Bayesian model for computing expected counts rather than using unadjusted historical counts currently, is under development.

### *Alert Notification*

When an algorithm triggers an alert based on the above criteria, RODS sends e-mail and/or page alerts to its users. RODS uses an XML-based configuration file to define users' e-mail and pager addresses. The e-mail version of the alert includes a URL link to a graph of the time series that triggered the alarm with two comparison time series: total visits for the same time period and normalized counts.

### *User Interface*

RODS has a password-protected, encrypted Web site at which users can review health care registration and sales of OTC health care products on epidemic plots and maps. When a user logs in, RODS will check the user's profile and will display data only for his or her health department's jurisdiction. The interface comprises three screens—Main, Epiplot, and Mapplot.

The *main screen* alternates views automatically among each of the available data sources (currently health care registrations and OTC products in Pennsylvania and Utah and OTC sales only for other states). The view alternates every two minutes as shown in Figure 3. The *clinic visits view* shows daily total visits and seven daily syndromes for the past week. The *OTC data view* shows daily sales for five product categories and the total, also for the past week. Users also can set the view to a specific county in a state. If the *normalize* control box is checked, the counts in the time series being displayed will be divided by (normalized by) the total daily sales of OTC health care products or ED visits for the region.

The *Epiplot* screen provides a general epidemic plotting capability. The user can simultaneously view a mixture of different syndromes and OTC product categories for any geographic region (state, county, or zip code), and for any time interval. The user also can retrieve case details as shown in Figure 4. The *Get Cases* button queries the database for the admission date, age, zip code, and chief complaint (verbatim, not classified into syndrome category) of all patients in the time interval and typically is used to examine an anomalous density (spike) of cases. The *Download Data* button will download data as a compressed comma separated file for further analyses.

The *Mapplot* screen is an interface to ArcIMS, an Internet-enabled GIS product developed by Environmental Systems Research Institute, Inc. Mapplot colors zip code regions to indicate the proportion of patients presenting with a particular syndrome. The GIS server also can overlay state boundaries, county boundaries, water bodies, hospital locations, landmarks, streets, and highways on the public health data as shown in Figure 5. Similar to Epiplot, Mapplot also can display case details for a user-selected zip code.

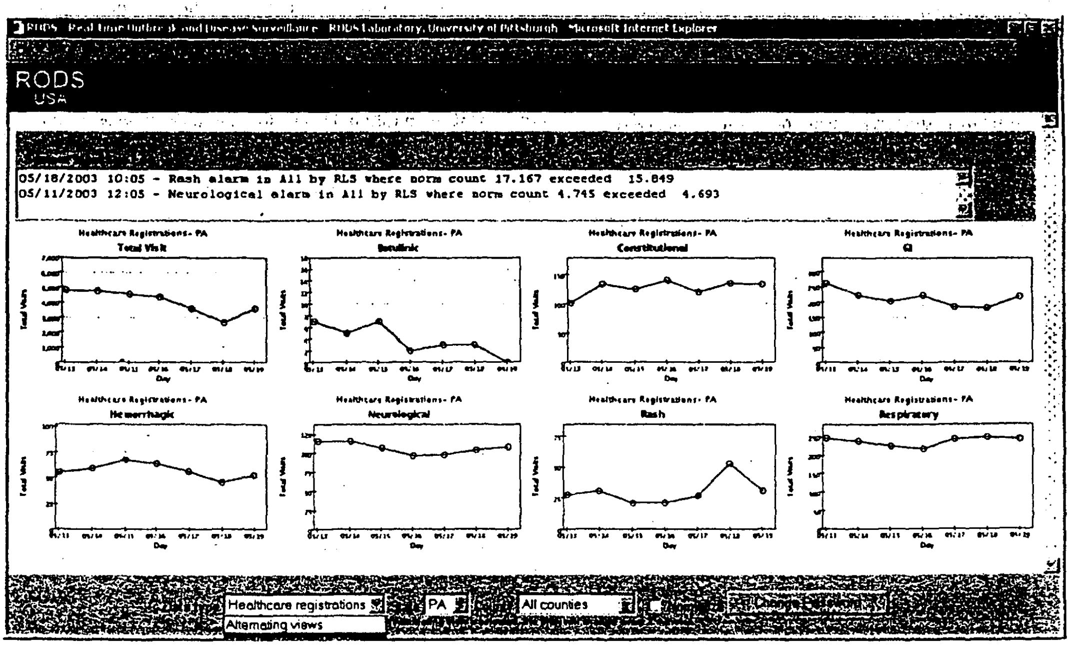
## Fault Tolerance

RODS has been in operation for four years and, like most production systems, has acquired many fault-tolerant features. For example, at the software level, HL7 listeners continue to receive messages and temporarily store the messages when the database is off-line. A data manager program runs every ten minutes and, on finding such a cache, it loads the unstored messages to the database when the database is back on-line. In addition, the data manager program monitors and restarts HL7 listeners as necessary. The database uses "archive log" mode to log every transaction to ensure that the database can recover from a system failure.

The hardware architecture also is fault tolerant. All servers have dual power supplies and dual network cards. All hard drives use Redundant Arrays of Inexpensive Disk configurations. In addition to dual power supplies, all machines are connected to an uninterrupted power supply that is capable of sending an e-mail alert to the RODS administrator when the main power is down.

## Health System Resident Component

An important component of RODS that currently is used only at the UPMC Health System in Pittsburgh is the Health System Resident Component (HSRC). The HSRC is located within the firewall of a health system and connects directly to the HL7 message router. The HSRC currently receives a diverse set of clinical data from the HL7 message router including culture results, radiology reports, and dictated



```
05/18/2003 10:05 - Rash alarm in All by RLS where norm count 17.167 exceeded  15.849
05/11/2003 12:05 - Neurological alarm in All by RLS where norm count 4.745 exceeded  4.693
```

RODS Real Time Outbreak and Disease Surveillance RODS Laboratory, University of Pittsburgh Microsoft Internet Explorer
RODS
USA
Healthcare Registrations- PA
Total Visit
Healthcare Registrations- PA
Botulinic
Healthcare Registrations- PA
Constitutional
Healthcare Registrations- PA
GI
Healthcare Registrations- PA
Hemorrhagic
Healthcare Registrations- PA
Neurological
Healthcare Registrations- PA
Rash
Healthcare Registrations- PA
Respiratory
Total Visits
Day
Healthcare registrations
PA
All counties
Alternating views
Change

**Figure 3.** *Health care registrations view* in the Main screen of RODS. The Main screen alternates views every 2 minutes among data types available in the public health jurisdiction. The figure shows eight plots of health care registration data—total visits, botulinic, constitutional, gastrointestinal (GI), hemorrhagic, neurological, rash, and respiratory. After 2 minutes, over-the-counter data will be displayed. The Main screen can be used as a "situation room" display.

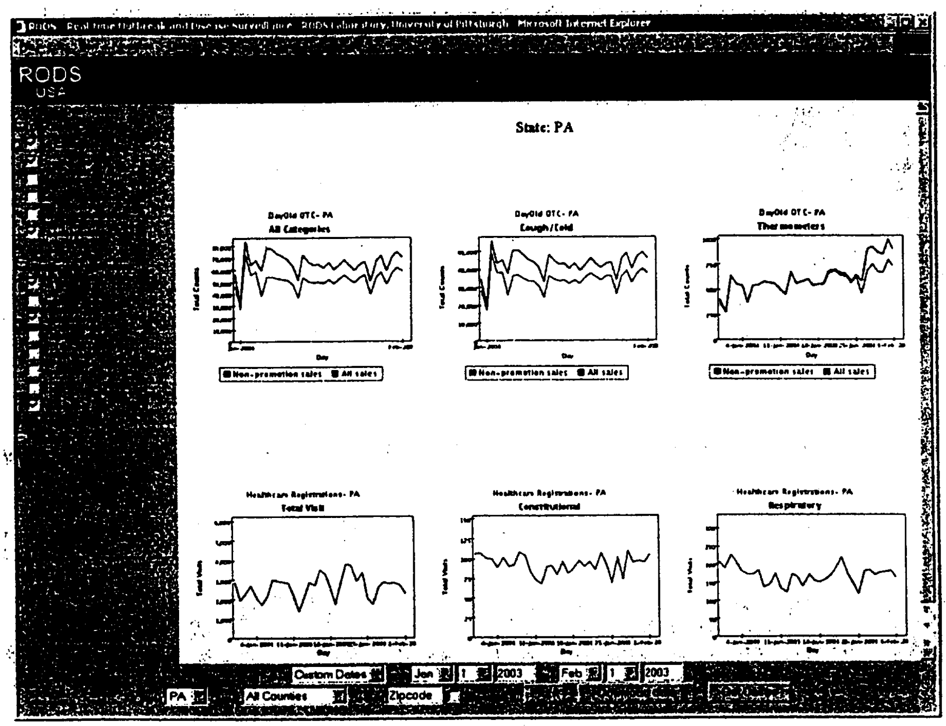


**Figure 4.** *Epiplot* screen of RODS. The six graphs are user-selected time-series plots of emergency department visits and sales of over-the-counter (OTC) products in Pennsylvania—OTC all categories, OTC cough/cold, OTC thermometers, clinic total visits, clinic constitutional, and clinic respiratory—between January 1, 2003, and February 1, 2003. Users can view a mixture of different syndromes and OTC product categories for any geographic region (state, county, or zip code), and for any time interval. Users can select types of data from the pick-list on the left of the screen. The *Download data* button retrieves raw count data for the selected graphs to a compressed comma-separated file. The *Get Cases* button shows a list of records containing chief complaint, age in decile, gender, and patient home zip code within the specified time interval and the geographic region. The lower, red line in the OTC plots represents nonpromoted sales.

emergency room notes. Its purpose is to provide additional public health surveillance functions that would not be possible if it were located outside of the firewall due to restrictions on the release of identifiable clinical data. The HSRC uses patient identifiers to link laboratory and radiology information to perform case detection. In the past, we have used HSRC to monitor for patients with both a gram-positive rod in a preliminary microbiology culture report and "mediastinal widening" in a radiology report. The HSRC is a case detector in a distributed outbreak detection system that is capable of achieving much higher specificity of patient diagnostic categorization through access to more information.

HSRC also removes identifiable information before transmitting data to the RODS system, a function provided by the health system's message router in other hospitals that connect to RODS.

The HSRC at UPMC Health System functions as an electronic laboratory reporting system, although the state and local health departments are not yet ready to receive real-time messaging from the system. Currently, it sends email alerts to the director of the laboratory and hospital infection control group about positive cultures for organisms that are required to be reported to public health in the state of Pennsylvania.[30] It also sends messages to hospital infection control when it detects organisms that cause nosocomial infections. These organisms include *Clostridium difficile*, methicillin-resistant *Staphylococcus aureus*, and vancomycin-resistant *Enterococcus*.

We have been able in HSRC to prototype one additional feature, which is a "look-back" function that facilitates very rapid outbreak investigations by providing access to electronic medical records to public health investigators as shown in Figure 6. This feature requires a token that can be passed to a hospital information system that can uniquely identify a patient, and the reason we have prototyped this feature in the HSRC and not in RODS is simply that HSRC runs within the firewall so an unencrypted token can be used. The look-back is accomplished as follows: when a public health user identifies an anonymous patient record of interest (e.g., one of 20 patients with diarrhea today from one zip code), HSRC calls the UPMC Health System Web-based electronic medical record system and passes it the patient identifier. UPMC Health System then requests the user to log in using the UPMC-issued password before providing access to the record

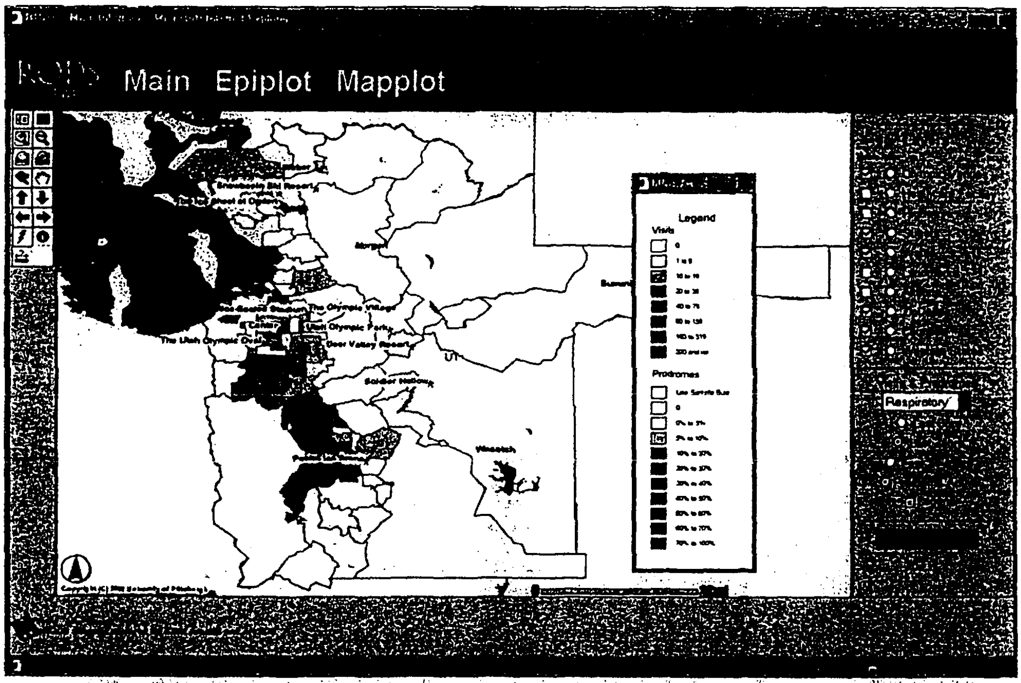


**Figure 5.** *Mapplot* screen of RODS shows spatial distribution of respiratory cases in part of Utah. Olympic venues are labeled.

directly from its own secure Web site. This approach is not intended to be implemented in HSRC, but rather in the RODS system outside of the firewall of a health system. It is intended to use encrypted identifiers that the health system would decrypt to retrieve the correct record. The HSRC could provide the encryption-decryption service or it could be provided by another data system in the hospital. We estimate that the prevalence of health systems that have Web-based results review in the United States is 30% to 50% and growing so that this approach could very quickly improve the efficiency of outbreak investigations.

## Current Status

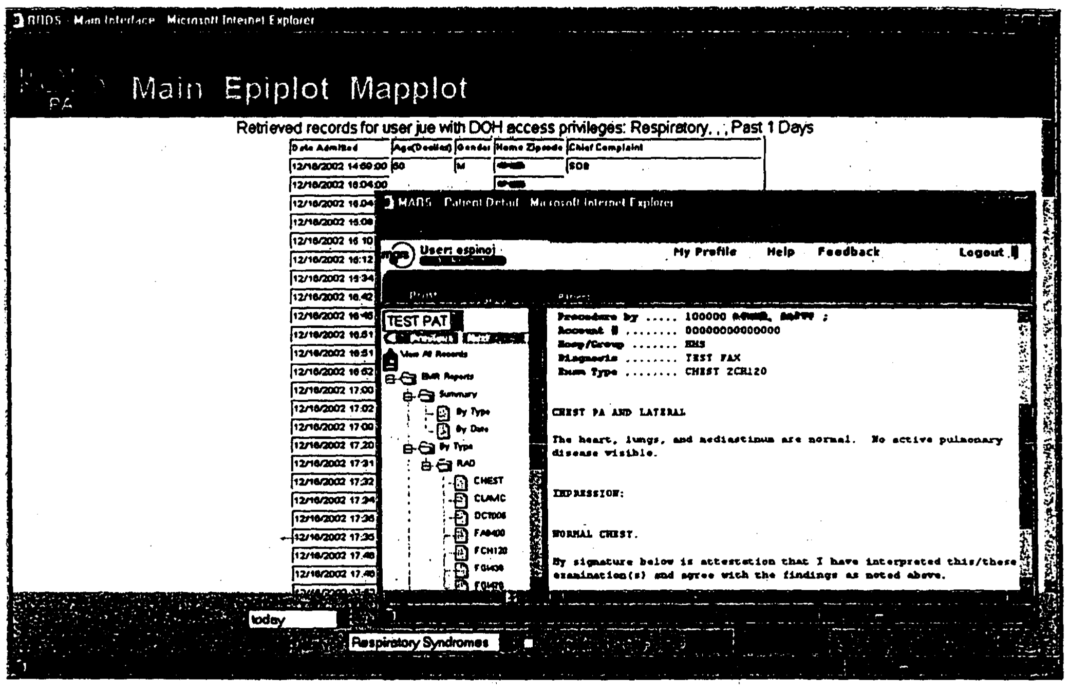
RODS has been in operation in Pennsylvania since 1999 and in Utah since January 2002. In Utah, RODS receives data from two health systems: Intermountain Heath Care, including nine EDs and 18 acute care facilities, and the University of Utah Health Sciences Center, with one ED.[24] Together, these facilities serve about 70% of the population of Utah. In Pennsylvania, RODS receives data from 20 health systems comprising 38 hospitals. Two health systems (each with one hospital) send plain text files to RODS on a daily basis. In Pennsylvania, RODS covers 80% of ED visits in Allegheny County (population, 1.3 M) where Pittsburgh is located; 50% of visits in the 13-county Metropolitan Medical Response Area centered on Pittsburgh (population 3.0 M); and more than 70% coverage of three other counties, including Dauphin County where Harrisburg, the capital of Pennsylvania, is located. The Commonwealth of Pennsylvania is funding a large project to connect the remaining hospitals in the Commonwealth with RODS over the next two years (approximately an additional 170 hospitals).

In December 2002, the RODS laboratory released version 1.1 of the RODS software to the public. The release includes all of the components necessary to deploy RODS for clinic visits surveillance. RODS is free for noncommercial use and can be downloaded at <http://www.health.pitt.edu/rods/sw/>. Although the software has been downloaded in excess of 170 times, we are aware of only a few successful efforts at deployment. These kinds of systems require network engineers, Oracle database administrators, and interface engineers, and very few health departments have access to that skills set.

For these reasons, we have moved to an application service provider model for dissemination in which we encourage state and local health departments to form coalitions to support shared services. We also have been fortunate to have sufficient grant funding from the Commonwealth of Pennsylvania to be able to support these services on an interim basis while sustainable funding models evolve.

## Discussion

Our original design objectives for RODS were real-time collection of data with sufficient geographic coverage and sampling density to provide early syndromic warning of a large-scale aerosol release of anthrax. Although we have not achieved all of our initial design objectives, progress has been substantial. The research identified two types of data—free-text chief complaints and sales of OTC health care prod-



RODS · Main Interface · Microsoft Internet Explorer

Main Epiplot Mapplot

Retrieved records for user jue with DOH access privileges: Respiratory, , , Past 1 Days

| Date Admitted | Age(Decile) | Gender | Home Zipcode | Chief Complaint |
|---|---|---|---|---|
| 12/18/2002 14:00:00 | 60 | M | [redacted] | SOB |
| 12/18/2002 16:04:00 | | | [redacted] | |
| 12/18/2002 16:04 | | | | |
| 12/18/2002 15:08 | | | | |
| 12/18/2002 16:10 | | | | |
| 12/18/2002 16:12 | | | | |
| 12/18/2002 15:34 | | | | |
| 12/18/2002 16:42 | | | | |
| 12/18/2002 16:46 | | | | |
| 12/18/2002 16:51 | | | | |
| 12/18/2002 16:51 | | | | |
| 12/18/2002 16:52 | | | | |
| 12/18/2002 17:00 | | | | |
| 12/18/2002 17:02 | | | | |
| 12/18/2002 17:00 | | | | |
| 12/18/2002 17:20 | | | | |
| 12/18/2002 17:31 | | | | |
| 12/18/2002 17:32 | | | | |
| 12/18/2002 17:34 | | | | |
| 12/18/2002 17:36 | | | | |
| 12/18/2002 17:35 | | | | |
| 12/18/2002 17:46 | | | | |
| 12/18/2002 17:46 | | | | |

MARS · Patient Detail · Microsoft Internet Explorer

User: espinoj

My Profile Help Feedback Logout

TEST PAT

View All Records

```
Procedure by ..... 100000 [redacted] ;
Account # ........ 00000000000000
Roop/Group ....... EMS
Diagnosis ........ TEST FAX
Exam Type ........ CHEST ZCH120


CHEST PA AND LATERAL

The heart, lungs, and mediastinum are normal.  No active pulmonary
disease visible.


IMPRESSION:


NORMAL CHEST.

By signature below is attestation that I have interpreted this/these
examination(s) and agree with the findings as noted above.
```

**Figure 6.** Look-back function of RODS. The user has selected one patient to investigate using the screen that is in the background and partly hidden by overlap. RODS has logged the user into the results-review function of an electronic medical record and requested that patient's chart, which is shown on the screen in the foreground.

ucts—that can be obtained in real time or near real time at sampling levels of 70% or higher for most of the United States. These results were obtained through large-scale deployments of RODS in Pennsylvania and Utah and through building the National Retail Data Monitor described in the accompanying article in this issue of *JAMIA*. The deployments also provided insights about organizational and technical success factors that would inform an effort to scale the project nationally.

The project established the importance of HL7 message routers (also known as integration engines) for public health surveillance. HL7 message routers are a mature, highly prevalent technology in health care. We demonstrated that free-text triage chief complaints can be obtained in real time from most U.S. hospitals through message routers and that these data represent early syndromal information about disease. Many other clinical data of value to public health are transmitted using the HL7 standard (e.g., orders for diagnostic tests, especially microbiology tests, reports of chest radiographs, medications, and test results) and can be integrated into RODS or other surveillance systems capable of receiving HL7 messages.

As a result of our efforts to disseminate this technology by giving it away, we have learned that most health departments do not have the technical resources to build and maintain real-time electronic disease surveillance systems. Our application service provider model has been much more successful, and we now recommend that states form coalitions to share the costs of such services.

The project very early identified the need for a computing component to reside within the firewall of a health system, connected to the hospital's HL7 message router. This component would function as a case detector in a distributed public health surveillance scheme linking laboratory and radiology data to increase the specificity of case detection. It has proven very difficult to disseminate this technology, perhaps due to the complexity of the idea. Nevertheless, the threat of bioterrorism has created a need for such technology, and this approach, or something with equivalent function, must be deployed.

Adherence to NEDSS architectural standards was an early design objective that we have met. RODS 1.5 closely follows NEDSS architectural, software, messaging, and data specifications. Our success is a strong validation of those standards. We will gain further understanding of the standards as we attempt to use RODS components including HL7 listeners, natural language parsers, message parsers, databases, user interfaces, notification subsystems, and detection algorithms with other NEDSS compliant systems. An ongoing project will use RODS to collect chief complaints and integrate them into the Utah Department of Health's planned NEDSS system.

We have demonstrated the ability to rapidly deploy RODS in a special event with the added advantage that the system persisted after the event. This experience suggests strongly that RODS or similar systems be considered an alternative to drop-in surveillance.

Our future plans are to meet our initial design objective to develop early-warning capability for a large, outdoor release of anthrax, especially ensuring that the data and analysis produced by RODS are reviewed by public health. This goal will require improvements in the interfaces and the detection algorithms to reduce false alarms and to vastly improve the efficiency with which anomalies are evaluated by use of multiple types of data, better interfaces, and implementation of the look-back function. We would like to enlarge as quickly as possible the application service provider to include more states and more types of clinical data so that states will be in a position to prospectively evaluate the detection performance from different types of data on naturally occurring outbreaks.

Our long-term goals are to add additional disease scenarios to the design objectives such as detection of in-building anthrax release, vector-borne disease, food-borne disease, and a communicable disease such as severe acute respiratory syndrome (SARS).

## Conclusion

RODS is a NEDSS-compliant public health surveillance system that focuses on real-time collection and analysis of data routinely collected for other purposes. RODS is deployed in two states and was installed quickly in seven weeks for the 2002 Olympics. Our experience demonstrates the feasibility of such a surveillance system and the challenges involved.

Outbreaks, emerging infections, and bioterrorism have become serious threats. It is our hope that the front-line of public health workers, astute citizens, and health care workers will detect outbreaks early enough so that systems such as RODS are not needed. However, timely outbreak detection is too important to be left to human detection alone. The notion that public health can operate optimally without timely electronic information is as unwise as having commercial airline pilots taking off without weather forecasts and radar.

*References*

1. Kaufmann A, Meltzer M, Schmid G. The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? Emerg Infect Dis. 1997;3(2):83–94.
2. Guillemin J. Anthrax: the investigation of a deadly outbreak. N Engl J Med. 2000;343:1198.
3. Siegrist DW. The threat of biological attack: why concern now? Emerg Infect Dis. 1999;5:505–8.
4. Roush S, Birkhead G, Koo D, Cobb A, Fleming D. Mandatory reporting of diseases and conditions by health care professionals and laboratories. JAMA. 1999;282:164–70.
5. Ashford DA, Kaiser RM, Bales ME, et al. Planning against biological terrorism: lessons from outbreak investigations. Emerg Infect Dis. 2003;9:515–9.
6. Silver J. Local doctors fail their test on diagnosing germ terrorism. Pittsburgh Post-Gazette. 2000;February 13. Available at: http://www.post-gazette.com/healthscience/2000213biowar3.asp. Accessed July 13, 2003.
7. Aston C. Biological warfare canaries [biological attack detection]. IEEE Spectrum. 2001;38(10):35–40.
8. Ackelsberg J, Layton M. Update #5: Terrorist Attack at the World Trade Center in New York City: Medical and Public Health Issues [online] 2001. <http://www.nyc.gov/html/doh/html/cd/wtcf.html>. Accessed May 16, 2003.
9. Wagner MM, Aryel R, Dato V. Availability and Comparative Value of Data Elements Required for an Effective Bioterrorism Detection System. Washington, DC: Agency for Healthcare Research and Quality, 2001.
10. Lewis MD, Pavlin JA, Mansfield JL, et al. Disease outbreak detection system using syndromic data in the greater Washington DC area. Am J Prev Med. 2002;23:180–6.
11. Lober WB, Thomas Karras B, Wagner MM, et al. Roundtable on bioterrorism detection: information system-based surveillance. J Am Med Inform Assoc. 2002;9:105–15.
12. Tsui F-C, Espino JU, Wagner MM, et al. Data, network, and application: technical description of the Utah RODS Winter Olympic Biosurveillance System. Proc AMIA Symp. 2002:815–9.
13. Paulson T. Region alert to bioterror, but health-care system underfunded [online] 2001. <http://seattlepi.nwsource.com/local/40829_bio29.shtml>. Accessed March 6, 2002.
14. Pueschel M. DARPA System Tracked Inauguration For Attack [online] 2001. <http://www.usmedicine.com/article.cfm?articleI D=172&issueID=25>. Accessed March 6, 2002.
15. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine. J Public Health Manag Pract. 2001;7(6):43–50.
16. NEDSS systems architecture. April 15, 2001. Available at: http://www.cdc.gov/nedss/nedssarchitecture/nedsssysarch2.0.pdf. Accessed July 13, 2003.
17. Thacker S, Berkelman R. Public health surveillance in the United States. Epidemiol Rev. 1988;10:164–90.
18. DoD-GEIS. Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE) [online] 2003. <http://www.geis.ha.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp>. Accessed May 16, 2003.
19. Schafer, K. LEADERS (Lightweight Epidemiology Advanced Detection & Emergency Response System) [online] 2001. <http://www.tricare.osd.mil/conferences/2001/agenda.cfm>. Accessed May 10, 2001.
20. Zelicoff A, Brillman J, Forslund D, et al. The Rapid Syndrome Validation Project (RSVP) [online] 2001. <http://www.cmc.sandia.gov/bio/rsvp/SAND%20No.pdf>. Accessed May 17, 2003.
21. Tsui F-C, Wagner MM, Dato V, Chang C-CH. Value of ICD-9-coded chief complaints for detection of epidemics. Proc AMIA Symp. 2001:711–5.
22. Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. Proc AMIA Symp. 2002:345–9.
23. Wagner MM, Robinson JM, Tsui F-C, Espino JU, Hogan WR. Design of a national retail data monitor for public health surveillance. J Am Med Inform. 2003;10:409–18.
24. Gesteland PH, Gardner RM, Tsui F-C, et al. Automated syndromic surveillance for the 2002 Winter Olympics. J Am Med Inform. 2003;10:(in press).
25. Liu Z, Tsui F-C, Zeng X. Cache table design for disease surveillance system. Proc AMIA Symp. 2002:1086.
26. Java 2 Platform Enterprise Edition (J2EE) [online] 2003. <http://java.sun.com/j2ee/>. Accessed February 10, 2003.
27. Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. Proc 16th Int FLAIRS Conference. 2003:412–6.
28. Orfanidis SJ. Optimum Signal Processing (ed 2). New York: McGraw-Hill, 1988.
29. Wong W, Moore A, Cooper G, Wagner M. Rule-based anomaly pattern detection for detecting disease outbreaks. Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI); 2002.
30. Panackal AA, M'ikanatha NM, Tsui F-C, et al. Automatic electronic laboratory-based reporting of notifiable infectious diseases. Emerg Infect Dis. 2001;8:685–91.