

# Data and Information Management in Public Health

Adrienne S. Ettinger, Sc.D., M.P.H.

Environmental Public Health Tracking  
Methods Course

July 2004

# Outline

- **Information Management in Public Health**
  - Information
  - Infrastructure
  - Informatics
- **Database Design**
  - Relational Model
  - Data Linkage
- **Data Mining and Data Warehousing**



# Information Management – why?

- **Data needs**
  - Need for good record-keeping and documentation
  - Need for program evaluation
  - Need high quality data to support valid inference
- **Data vs. Information**
  - Public health tradition of generating data
  - Staff time and skills not being spent on analysis
  - Possibility of automating analyses

J2

Perhaps need to clarify the difference between data and information

JHerbstman, 2/2/2005

# Information Integration

- **Lack of existing data standards**
- **Incompatible systems**
- **Paper systems**
- **Categorical stand-alone systems**
- **Inability to identify (link) individuals being served in multiple systems**
- **Integration of individual multiple records**

# Rates of IT Failure: High

- **16.2% were “project successful”**  
(software projects that are completed on-time and on-budget among American companies and governments)
- **52.7% were “project challenged”**  
(they were completed and operational but over-budget, over the time estimate, and offers fewer features and functions than originally scheduled)
- **31.1% were “project impaired”**  
(cancelled)

Source: *Charting the Seas of Information Technology*  
The Standish Group 1994

J1

What is the context for this slide? What aspect of IT?

JHerbstman, 2/2/2005



# Barriers to IT in Public Health

**1. Information**

**2. Infrastructure**

**3. Informatics**



# 1. Information

- **Surveillance data**
  - Only 15-20% of reportable cases reported
  - Delays of days to weeks
  - Not typically in electronic form
- **Other relevant data not electronically available**
  - Environment, injury, etc.
  - Guidelines
  - Contacts
  - Training materials

# Information in Progress

- **NEDSS = National Electronic Disease Surveillance System**
  - Architectural elements
  - Public health conceptual data model
- **Knowledge management**
  - **Preventioneffects.net**
  - **Encoding of clinical guidelines**
    - Disseminate
    - Point-of-care reminders

J3

Maybe an intro bullet point about how NEDSS and other efforts are being made to remedy some of the problems about public health 'Information'

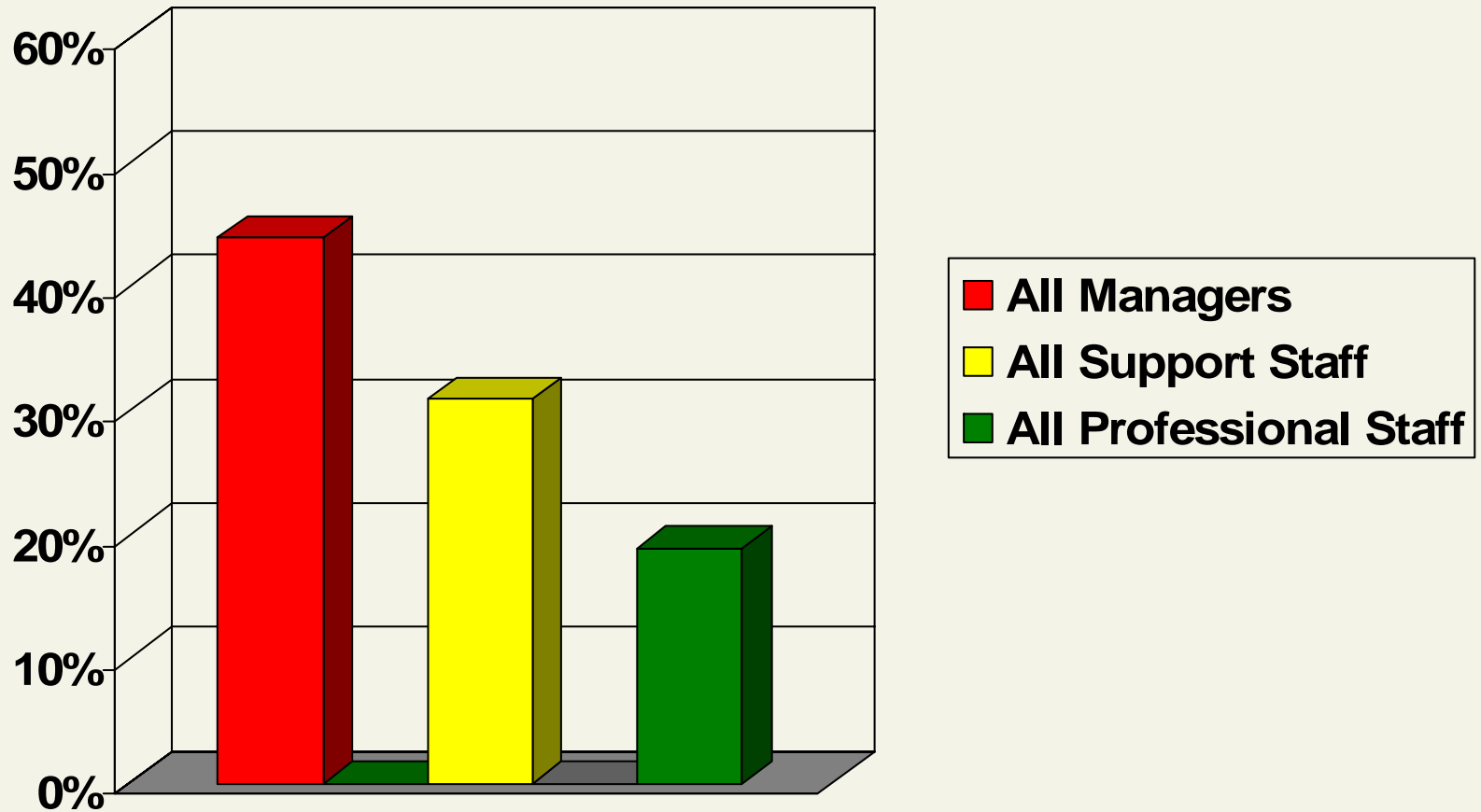
JHerbstman, 2/2/2005

## 2. Infrastructure

- **Information technology**
  - Only 48.9% of local health departments have high-speed continuous internet connections (NACCHO, 1999)
- **Workforce**
  - 83% of local health departments indicate that computer training is a key need (NACCHO, 1996)

# Desktop Web Access Among Minnesota Public Health Staff

by Job Class, February 2000



Source: Minnesota Health Alert Network (HAN) Project, 2000

# Infrastructure Development

- **INPHO = Information Network for Public Health Officials (state)**
  - Ending in FY2001
- **HAN = Health Alert Network (local)**
- **Frist-Kennedy authorization**
  - Infrastructure standards/ assessment
  - Preparedness of public health system

**J4** Perhaps a bit more information describing these efforts. . .maybe Frist-Kennedy bullet ok  
JHerbstman, 2/2/2005

# 3. Informatics

- **The systematic application of computer & information science and technology to public health practice, research, and learning...**



J6

need to integrate this slide and the next slide (12 and 13)

JHerbstman, 2/2/2005

# 3. Informatics

- **The systematic application of computer & information science and technology to public health practice, research, and learning...**through integrated information resource management planning; assembling and managing teams with diverse skill sets; managing tasks to complete projects, etc.

# Management Skills

- **IT projects expensive and high risk**
- **Interdisciplinary teams required**
- **New skills needed by public health managers**

J7 not sure the point of this slide. . .perhaps can be combined with the next one?  
JHerbstman, 2/2/2005

# Specific roles - public health managers

- **Specify requirements (minimum)**
- **Facilitate integrated, coordinated IT development (through advice, leadership)**
- **Manage specific IT projects; assemble and manage development teams**
- **Translate program vision for technical staff, and vice versa**
- **Appropriately procure IT products & services**
- **Resolve inevitable tensions**

# Informatics in Public Health

- **Information Access**
  - Databases
  - Knowledge management
- **Information Systems**
  - Effective management
  - Improved productivity
- **Surveillance integrated with EMR**
- **Feedback to providers**

# Why gather data?

## **Determine the magnitude of the problem**

- Data is the connection between the problem and how to solve the problem
- Describe what is known about the problem: person, place, time
- Place the problem in context
- Describe what already exists (prevention and intervention programs)
- Compare data to what should exist, identify gaps
- Identify populations or areas at high-risk
- Learn more about your community

# Why gather data?

## **Monitor trends over time**

- Provides a source of baseline information
- Progress can be measured against baseline benchmark



# Why gather data?

## **Provides information and a basis for decision-making**

- Set priorities
- Develop program based on current information
- Needs
- Resources
- Inform and convince decision-makers
- Need a roadmap to know where you are going and when you have arrived

# How to gather data?

- **Define the problem or question to be addressed**
  - Number of people affected
  - Place that is affected
  - Time period of analysis
- **Generate a hypothesis (educated guess) about the reason for the problem**
- **Identify sources of data to answer the question posed**
- **Define variables to measure problem or question**
- **Identify methods to be used to analyze data collected**

# What data to collect?

- **Types of Data**
  - Primary
  - Secondary
- **Levels of Data / Unit of Observation**
  - Individual-level data on persons or houses
  - Aggregate data at the community-level
- **Sources of Data**
  - Demographic characteristics (ex: vital statistics)
  - Geographic characteristics (ex: census data)
  - Socioeconomic Characteristics (ex: labor, education)
  - Health (ex: health department)
  - Environment (ex: state environmental protection)

# Hazard Data Sources

- ❖ Ambient Air Concentrations
- ❖ Air Emissions and Inspections
- ❖ Toxic Release Inventory
- ❖ Ground Water Sampling
- ❖ Drinking Water Databases
- ❖ Meteorology

# Exposure Data Sources

- ❖ Human Biomonitoring
- ❖ Personal Sampling
- ❖ Exposure Surrogates
  - Survey Data
  - Modeled Exposures

# Health Data Sources

- ❖ Notifiable diseases
- ❖ Laboratory specimens
- ❖ Vital records
- ❖ Sentinel surveillance
- ❖ Disease registries
- ❖ Periodic surveys
- ❖ Special studies
- ❖ Administrative data systems

# What data to collect?

## **Logistical considerations**

- Budget constraints
- Staffing time and expertise
- Available technology
- Planning for future updates
- Linkage and integrations of existing systems
- Security concerns
- User-friendliness
- Can the system be maintained

# Planning for Data Collection

- **Identify public health needs**
- **Identify users**
- **Identify purpose: Why build the system?**
- **Define objectives: How will the data be used?**
- **Establish case definitions and standards**
- **Integration with existing systems**
  - **functional**
  - **technical**



# Data Management Protocol

**Data management protocol defines:**

- Standard operating procedures**
- Data sources**
- Data collection procedures**
- Data file structure**
- Data dictionary/code book**
- Documentation and archiving**

# Evaluation of Data Sources

- Availability of data (format, access, approvals needed, cost)
- Comparability (across geographic areas)
- Coverage (local, state, national; missing data)
- Relevance for tracking (timeliness, etc.)
- Misclassification
- Ability to control confounding, individual level data
- Size, complexity, and format of data files (technology)

# Additional Considerations

- Legal requirements
- Confidentiality & security
- Analysis plan
  - Who
  - Table shells
  - Statistics
  - Periodicity
- Dissemination plan

# Database

- An organized collection of information (nowadays almost invariably electronic).
- In relational databases, the **table** is a fundamental building block.
- A database consists of one or more tables, which are **related** (conceptually linked) to each other.

# Table

- A structure that consists of **rows** and **columns**.
- The rows are also called records, the columns are also called **fields**.
- Example - a table of Students will have the fields:
  - Social Security Number
  - First name
  - Last name
  - Date of birth, etc.

There will be one row (record) for each student.

# Types of Data Configuration

- **Wide** (one record per person)
  - One line for every individual (name, date of birth, gender, race...)
- **Long** (many records per person)
  - Multiple lines for every individual
    - **Fixed** (visit 1, visit 2, visit 3...)
    - **Variable** (prescription drug utilization, number of diagnoses per hospitalization, number of procedures per visit...)

# Data Linkage

- ❖ “Linkage” is defined as the physical integration of different databases resulting from a merge that utilizes a common variable
- ❖ Integration of health surveillance and environmental monitoring systems for hazards and exposures

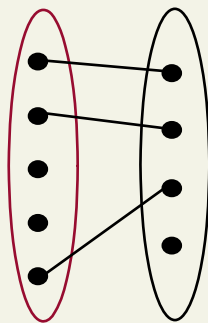
# Key field(s)

- A combination of one or more common variables (fields) that are used for indexed search whose value **uniquely** identifies a record in a table
- Therefore, no two records in a table can have the same key value.

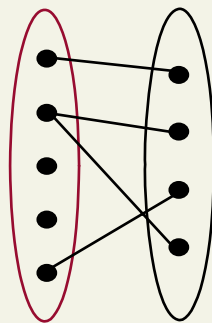


# Entities and Relationships

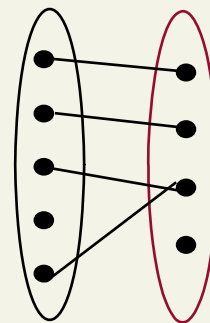
- When planning a database, one needs to identify **Entities** (the things about which we want to capture information) and the **Relationships** between them.
- Relationships between entities are **one-to-one**, **one-to-many**, or **many-to-many**.



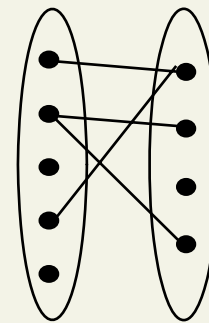
1-to-1



1-to Many



Many-to-1



Many-to-  
Many

# Relational Database

- Store data in tables
- Variables are grouped in logical units
  - By data source
  - By visit or interaction with system
  - By type of data (i.e. laboratory test)
- Normalize the tables
- Make prudent choice of primary key(s)
- It implies "logical" (proper) design of a database with minimal redundancy of data.

# Data Dictionary/Code Book

- Define and name the variables
- Data attribute, format, and range of permissible values
- Range and logic checks performed
- Coding scheme
  - Use standard coding scheme
  - Be consistent
  - Anticipate missing values

# Metadata

- “Data that describes other data”
- **Technical vs. Descriptive**
  - *Process-related* or *technical* metadata supports *software* efforts
  - *Descriptive* metadata, which supports users concerned with the software’s *application domain/s* (e.g., medicine, business).

# Enterprise Architecture

- The guiding structure and integrating framework for the design and development of information systems (IS)
- Encompasses broad decisions that must be made by an organization as it creates its organizational information support system

# Variable Attributes

- Number
  - Integer (whole), real (decimal)
  - Leading zeros
- Character/alphanumeric/text/string
- Logical value (yes/no, male/female)
- Date/time
  - different formats (MMDDYYYY, MMDDYY)
- Missing values
  - “special” missing

# Data Coding Standards

- Diagnosis codes (ICD-10)
- Medical procedures codes (CPT)
- National drug codes (U.S. FDA)
- Logical Observations Identifier Names and Codes (LONIC)
- Systematized Nomenclature of Medicine (SNOMED)
- Health Level 7 standard (HL-7)

# From data to analytic files

- **Raw data stored in data files should not be altered**
- **Quality “cleaning” of data**
  - **Range checks**
  - **Consistency**
- **Derived variables**
- **Merged files or variables from other files**



# Data Linkage

- A unique identifier is needed to link data from different sources

# Common Problems

- Duplicate records
- Merging of data files
  - 1-to-1 merge
  - 1-to-many merge
- Errors in programming logic for derived variables
- Inadequate documentation
- De-identification of records
- Version control
  - protocols, computer programs and reports

# Historical Perspectives

- **Hierarchical Databases** (mid 60s)
- **Network Databases** (late 60s)
- **Relational Databases** (late 60s to present)
- **Object-Oriented Databases and Object-Relational Databases** (late 80s to present)

J8

I'm not sure it is clear what these are or how they are different

JHerbstman, 2/2/2005

# Why Study the Relational Model?

- Most widely used model.
  - Vendors: IBM, Informix, Microsoft, Oracle, Sybase, etc.
- “Legacy systems” in older models
  - e.g., IBM’s IMS (hierarchical model)
- Recent competitor: object-oriented model
  - ObjectStore, Versant, Ontos, O2
  - A synthesis emerging: *object-relational model*
    - Informix UDS, UniSQL, Oracle, DB2

# SQL and file manipulation

- **Structured Query Language (SQL)**
  - Implemented in relational database management systems
- **Frequently used SQL commands**
  - Select variable(s)
  - Combine tables (merge)
  - Apply selection criteria (view, query)

# Data Mining

- The process of secondary data analysis of large databases aimed at finding suspected relationships which are of interest or value to the database owners.
  - Hand DJ. *Am Statistician* 1998; 52:112-8.
- Also known as: “Knowledge discovery”
- Keeping a watchful eye for unsuspected relationships by evaluating large datasets with many diseases and many variables of potential interest without a specific hypothesis

# Data Mining: Issues

- **No a priori hypothesis**
- **No pre-specified model form**
- **Multiple comparisons**
- **Expected counts**
- **Granularity**
- **Data mining tools create analytical models that are predictive, descriptive or both.**



# Data Warehousing

- The act of gathering data from distributed locations in a single store, usually in some aggregated form for further analysis.
- A data warehouse is a collection of data gathered and organized so that it can easily be analyzed, extracted, synthesized, and otherwise be used for the purposes of further understanding the data.
- It may be contrasted with data that is gathered to meet immediate objectives.

# Information and Data Systems

## **Challenges**

- Electronic communication gaps & fragmentation
- Many disparate systems
- Slow adoption of standards
- Technology just arriving on scene for many agencies
- Lack of financial resources

# Competing agendas

- **Build simple systems  $\Leftrightarrow$  address complex problems**
- **Solve immediate problem  $\Leftrightarrow$  build an integrated IT environment**
- **Program specialists  $\Leftrightarrow$  IT specialists**
- **“Get it done”  $\Leftrightarrow$  “Do it right”**
- **Build application today  $\Leftrightarrow$  Build foundation for tomorrow**

# Data Linkage – the details

- Data collected for different purposes
- Level of specificity or reporting may not be sufficient (aggregate data)
- Access or permission to use data difficult to obtain, cost or fees associated with use
- Information needed to conduct an epidemiologic study can vary greatly from what is needed for surveillance
- Inadequate variable(s) for indexing
- Methodological limitations



# Critical Questions

- ❖ Is it possible to “retro-fit” existing data systems for environmental public health tracking?
- ❖ How can we use the lessons learned to move forward with recommendations for new data collection for tracking?