
Bioinformatics

ROBERT E. HOYT
INDRA NEIL SARKAR

Learning Objectives

After reading this chapter the reader should be able to:

- Define bioinformatics, translational bioinformatics and other bioinformatics-related terms
- State the importance of bioinformatics in future medical treatments and prevention
- Describe the Human Genome Project and its many important implications
- List private and governmental bioinformatics databases and projects
- Enumerate several bioinformatics projects that involve electronic health records
- Describe the application of bioinformatics in genetic profiling of individuals and large populations

Introduction

In this chapter we will discuss bioinformatics, the biomedical informatics sub-discipline that has gained increasing prominence in recent years thanks to initiatives such as the Human Genome Project, discussed in a later section. Bioinformatics can trace its formal beginning to about 30 years ago. However, in many ways bioinformatics has evolved independent of health informatics and thus has its own sets of definitions and background information.

Definitions

We begin with some common definitions and in the next section provide a short genomics primer.

- Bioinformatics, often times referred to as Computational Biology, is a general description of "the field of science in which biology, computer science and information technology merge to form a single discipline". Bioinformatics makes use of fundamental aspects of computer science (such as databases and artificial intelligence) to develop algorithms for facilitating the development and testing of biological hypotheses, such as: finding the genes of various organisms, predicting the structure and/or function of newly developed proteins, developing protein models and examine evolutionary relationships.^{2,3}
- Translational bioinformatics focuses on the "development of storage, analytic and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into

proactive, predictive, preventive and participatory health.⁷⁴ Simply put, translational bioinformatics is the specialization of bioinformatics for human health.

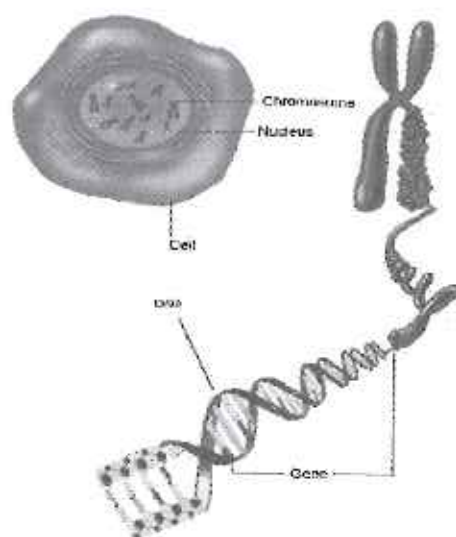
- Genomics is the field that analyzes genetic material from a species
- Proteomics is the study at the level of proteins (e.g., through gene expression)
- Pharmacogenomics is the study of genetic material in relationship with drug targets
- Metabolomics is the study of genes, proteins or metabolites
- Metagenomics is the analysis of genetic material derived from complete microbial communities harvested from natural environments⁷⁵
- Phenotype is the observable characteristic, structure, function and behavior of a living organism. Size and hair color could be examples. Phenotype is largely determined by the genotype
- Genotype is the genetic information that is often associated with phenotypes or regulation of biological function⁶

Genomic Primer

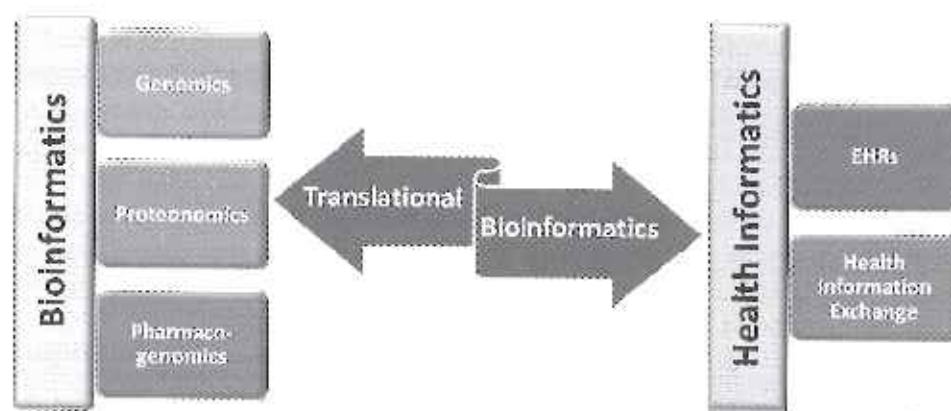
The human body has about 100 trillion cells and each one contains a complete set of genetic information (chromosomes) in the nucleus; exceptions are eggs, sperm and red blood cells. Humans have a pair of 23 chromosomes in each cell that includes an X and Y chromosome for males and two Xs for females. Offspring inherit one pair from each parent. Chromosomes are listed approximately by size with chromosome 1 being the largest and chromosome 22 the smallest. Organisms have differing numbers of chromosomes (e.g., our closest extant primate relatives, chimpanzees, have 24 pairs). Chromosomes consist of double twisted helices of deoxyribonucleic acid (DNA). DNA is composed of four sugar-based building blocks ("nucleotides": adenine [A], thymine [T], cytosine [C], and guanine [G]) that are generally found in pairs ("Watson-Crick" pairing: A-T, C-G). Genes are regions on chromosomes that encode instructions, which may result in proteins that then in turn enable biological functions. The process of decoding genes involves transcribing the DNA into ribonucleic acid (RNA) and then translation into amino acids that make up proteins (Figure 22.1). Collectively, the complete set of genes is referred to as a "genome." It is estimated that humans have between 20,000 and 30,000 genes and that genomes are about 99.9% the same between individuals. Variations in genomes between individuals are known as single nucleotide polymorphisms (SNPs) (pronounced "snips"). There are three types of alterations: single base-pair changes, insertions or deletions of nucleotides, and reshuffled DNA sequences. Although SNPs are common, their significance is complex and unpredictable.⁷⁻⁹

Importance of Bioinformatics

Besides diagnosing the 3,000 to 4,000 hereditary diseases that exist today, bioinformatics may be helpful to discover more targets for future drugs, develop personalized drugs based on genetic profiles and develop gene therapies to treat diseases with a strong genomic component, such as cancer. The most common way to achieve this is to use genetically altered viruses that carry human DNA. This approach, however, has not been definitely shown to work and has not been for general use by the FDA. Manipulation of genomes in other organisms, such as microbes, has shown promise for energy production ("bio-fuels"), environmental cleanup, industrial processing and waste reduction. Genetically engineered plants could also be made to be drought or disease resistant.

Figure 21.1: Genes (Courtesy of Nat. Inst. of General Medical Sciences)

This chapter will deal primarily with transformational bioinformatics (TBI), a relatively newly identified area of focus in bioinformatics that is largely focused on the study of data contained within exponentially growing genetic and clinical databases. A significant goal of TBI is to enable bi-directional crossing of the translational barrier between the research bench and the bed in the medical clinic. With growing genome-wide and population-based research data sets we are uncovering more genotype-phenotype associations that potentially can detect and treat diseases with a genetic component earlier. Such associations may also help create tailor made drugs for higher efficacy. Figure 22.2 demonstrates the bidirectional nature of data and information flow between bioinformatics and health informatics. We have seen the emergence of translational bioinformatics primarily due to the rapid advances in technology on both sides. In other words, a variety of advances in bioinformatics, such as faster and cheaper DNA sequencing, and more widespread adoption of electronic health records have made this possible.

Figure 22.2: Translational bioinformatics (Adapted from Sarkar et al¹⁰)

Pharmacogenomics is an excellent example of how translational bioinformatics can be used within the context of pharmaceutical development to utilize genomic information for better drug discovery and utilization. Drug companies are faced with the huge expense of drug development, the long road to producing a new drug and expiring patents. Drug failures are common and can be due to lack of clinical efficacy, side effects and commercial issues. Unfortunately, animal models are often times not adequate for the

development and evaluation of drugs for treating human conditions. It is thus the goal to use genetic information for:

- New indications for an old drug (drug repurposing)
- New targets for existing drugs (e.g., treatment of tongue cancer using RET inhibitors)
- Drugs to work better in certain patient groups (gender, age, race, ethnicity, etc.) with possible genetic variants
- Knowing ahead of time what drugs to avoid due to higher incidence of side effects that are genetically modulated
- Develop clinical decision support in electronic health records based on pharmacogenomics^{11,12}

Multiple projects are underway to integrate genetic and clinical data that will be discussed later in the chapter. We want to emphasize that burgeoning electronic health records (EHRs) and health information exchanges (HIEs), which are rapidly becoming ubiquitous, will contribute massive amounts of patient information (including demographic, laboratory, and clinical data). It is important to also note that in addition to genomic and clinical data, environmental data may offer valuable insights into the understanding and eventual treatment of disease.

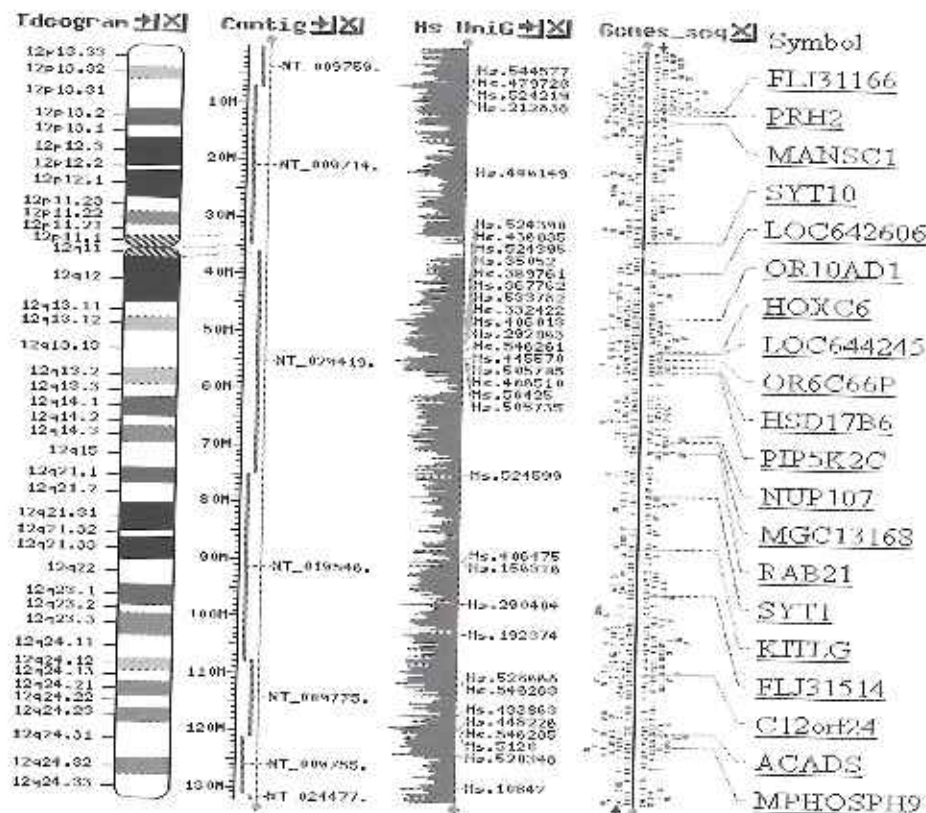
Bioinformatics Projects and Centers

The Human Genome Project (HGP)

One of the greatest accomplishments in medicine in the current era of science was the Human Genome Project. This international collaborative project, sponsored by the US Department of Energy and the National Institutes of Health, was started in 1990 and finished in 2003. In the process of acquiring the human genome (as a complete set of DNA sequence, encompassing all 23 chromosomes), genome sequences for a number of other key organisms ("model" organisms) were also acquired. These included the *Escherichia coli* bacterium, fruit fly (*Drosophila melanogaster*), and house mouse (*Mus musculus*). By mid-2007 about three million differences (SNPs) had been identified in human genomes. Appreciating the potential significant societal impact, the HGP also addressed the ethical, legal and social issues associated with the project. Since the completion of the HGP, attention is now more focused on the development of approaches to analyze and learn from volumes of data representing increasing numbers of individuals.¹³⁻¹⁵ These analyses include the annotation of information associated with disease onto chromosomes. Figure 22.3 displays the DNA sequencing of just chromosome number 12. Huge relational databases are necessary to store and retrieve this information. New technologies continue to emerge that reduce the necessity to sequence an entire human genome, such as DNA arrays (gene chips) that help speed the analysis and comparison of DNA fragments.¹⁶ The cost of the HGP was close to \$3 trillion; by 2010, a single gene chip can detect over a million variations in the base-pairs in a genome, in a few hours, costing only several hundred dollars.¹⁷ Even more exciting is the prospect that within a decade it is expected that the cost of an entire human genome will cost around \$1,000.

National Human Genome Research Institute (NHGRI)

NHGRI is an NIH institute that has many educational resources on their web site. Like other NIH institutes, they conduct and fund research within their intramural division, as well as support extramural research with external partners. Their health section has multiple resources for patients and healthcare professionals with particular emphasis on the Human Genome Project. The "Issues in Genetics" section covers important controversies in policy, legal and ethical issues in genetic research. They include a large glossary (200+) of genetics-related definitions, also available as a software app for the iPhone and iPad.¹⁷

Figure 22.3: Chromosome 12 (Courtesy of the National Library of Medicine)

Human Microbiome Project (HMP)

It is estimated that less than 0.01% of microbes on Earth have been cultured, characterized, and sequenced. As an exception, the complete genome for the common human parasite *Trichomonas vaginalis* was reported in 2007 in the journal *Science*.¹⁸ HMP is a NIH sponsored initiative that will study the myriad of organisms (oral, nasal, skin, gastrointestinal flora, etc.) that co-exist with humans and heretofore have been rarely studied. It will utilize metagenomics, as explained in the definitions section. As detailed on the HMP web site their goals are as follows:

- Determine whether individuals share a core human microbiome
- Understand whether changes in the human microbiome can be correlated with changes in human health
- Develop new technological and bioinformatic tools needed to support these goals
- Address the ethical, legal and social implications raised by human microbiome research























Human Variome Project

This Australian initiative began in 2006 with the goal to create systems and standards for storage, transmission and use of genetic variations to improve health. Rather than catalogue "normal" genomes they focus on the abnormalities that cause disease. Another aspect of their vision is to provide free public access to their databases.¹⁹

National Center for Biotechnology Information (NCBI)

The NCBI was created in 1988 and is part of the National Library of Medicine at the National Institutes of Health. It hosts thousands of databases associated with biomedicine (including the popular MEDLINE and GenBank databases) and thereby is considered one of the world's largest biomedical research centers. The NCBI provides access to sequences from over 100,000 organisms (via GenBank), including the complete genomes of over 1,000 (via NCBI Genome). Genomes represent both completely sequenced organisms and those for which sequencing is still in progress. Popular NCBI databases, which are linked by a common interface (Entrez), are listed in Figure 22.4.

Figure 22.4: NCBI Databases (Courtesy National Library of Medicine)

 Nucleotide: sequence database (includes GenBank)	 UniGene: gene-oriented clusters of transcript sequences
 Protein: sequence database	 CDD: conserved domain database
 Genome: whole genome sequences	 3D Domains: domains from Entrez Structure
 Structure: three-dimensional macromolecular structures	 Map: markers and mapping data
 Taxonomy: organisms in GenBank	 PopSet: population study data sets
 SNP: single nucleotide polymorphism	 GEO Profiles: expression and molecular abundance profiles
 Gene: gene-centered information	 GEO DataSets: experimental sets of GEO data
 HomoloGene: eukaryotic homology groups	 Cancer Chromosomes: cytogenetic databases
 PubChem Compound: unique small molecule chemical structures	 PubChem BioAssay: bioactivity screens of chemical substances
 PubChem Substance: deposited chemical substance records	 GENSAT: gene expression atlas of mouse central nervous system
 Genome Project: genome project information	 Probe: sequence-specific reagents

If you access the Genome project you can do a search for specific genes or proteins from different species. Figure 22.5 demonstrates the result of an Entrez Gene search for a tumor protein (TP53).

The NCBI site also provides access to BLAST (Basic Local Alignment Search Tool) that enables the identification of significantly related (based on a "expectation" value or "e-value") nucleotide or protein sequences from within the protein and nucleotide databases.²⁰

GenBank

This database was established in 1982 and is the NIH sequence database that is a collection of all publicly available DNA sequences. Along with EMBL (Europe) and DDBJ (Asia), GenBank is a member of the International Nucleotide Sequence Database Consortium (INSIG), which provides free access to sequence data from nearly anywhere with an internet connection. As of this writing, there are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions. Interestingly, many biological and medical journals now require submission of sequences to a database prior to publication, which can be done with NCBI tools such as BankIt.²¹

Figure 22.5: Entrez search for tumor protein (Courtesy National Library of Medicine)

NCBI Entrez Gene

Search: Gene for [] [Go] [Clear]

Limits: Previous: History: Clipboard: Details

Display: Full Report [] Page: 5 [] Send to: []

All 1 [] Current Only [] Genes (Genomes) [] SNP GeneView []

1: TP53 (tumor protein p53 (Li-Fraumeni syndrome)) [*Homo sapiens*]
 GeneID: 7157 Primary source: HGNC:11398 updated 30-Aug-2006

Summary

Official Symbol: TP53 and Name: tumor protein p53 (Li-Fraumeni syndrome) provided by HUGO Gene Nomenclature Committee
 See related: [HPEID:3,852](#), [CCDC191170](#)
 Gene type: protein coding
 Gene name: TP53
 Gene description: tumor protein p53 (Li-Fraumeni syndrome)
 RefSeq status: Reviewed
 Organism: *Homo sapiens*
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhina; Catarrhini; Hominoidea; Homo
 Gene aliases: p53, LFS1, TSP53
 Summary: Tumor protein p53, sometimes known as p51, is a central role in the regulation of cell cycle, specifically in the transition from

The Online Mendelian Inheritance in Man (OMIM)

This is another NCBI database of genetic data and human genetic disorders. It was originally developed and sponsored by Johns Hopkins University and Dr. Victor McKusick, a pioneer in genetic metabolic abnormalities. It includes an extensive reference section linked to PubMed that is continuously updated.²²

World Community Grid

This project was launched by IBM in 2004 and simply asked people to donate idle computer time. By 2007 over 500,000 computers were involved in creating a super-computer used in bioinformatics. Projects include Help defeat Cancer, Fight AIDS@Home, Genome Comparison and Human Proteome Folding projects. This grid promises to greatly expedite biomedical research by analyzing complex databases more rapidly as a result of this grid.²³

Pharmacogenomics Knowledge Base (PharmGKB)

This Stanford University based resource catalogues the relationships between genes, disease and drugs. There are sections on drugs, medical literature, variant genes, pathways, diseases and phenotypes that are searchable.²⁴

Framingham Heart Study SHaRe Genome-Wide Association Study

In 2007, the Framingham Heart Study began a new phase by genotyping 17,000+ subjects as part of the FHS SHaRe (SNP Health Association Resource) project. The SHaRe database is located at NCBI's dbGaP and will contain 550,000 SNPs and a vast array of phenotypical (combined characteristics of the genome and environment) information available in all three generations of FHS subjects. These will include measures of the major risk factors such as systolic blood pressure, total, LDL and HDL cholesterol, fasting glucose, and cigarette use, as well as anthropomorphic measures such as body mass index, biomarkers such as fibrinogen and C-reactive protein (CRP) and electrocardiography (EKG) measures such as the QT interval.²⁵

The Mayo Clinic Bipolar Disorder Biobank

Researchers at the Mayo clinic and other institutions are analyzing the genetic and clinical information on 2000 patients in their biobank to determine genetic aspects of bipolar disorder. It is hoped that data generated from this project will lead to earlier and better treatment of this mental health disorder.²⁶

Informatics for Integrating Biology and the Bedside (i2b2)

i2b2 is a National Institutes of Health National Center for Biomedical Computing located at Harvard Medical School. The Center has developed open source software that will enable investigators to mine existing clinical data for research. At this time there are 72 member institutions, including 12 that are international. The project was designed to allow users to query a system-wide de-identified repository for a set of patients meeting certain inclusion or exclusion criteria. On the web site, users can download client-software, client-server software and the source code.²⁷ The i2b2 infrastructure has been shown to be generalizable to multiple sites for a range of clinical conditions.²⁸

Cancer Biomedical Informatics Grid (CaBIG)

CaBIG is sponsored by the National Cancer Institute at the National Institutes of Health. The architecture is known as CaGrid and is an open source service oriented architecture (SOA). The infrastructure is designed to support the collection and analysis of data from disparate systems to promote biomedical research. The core software and associated tools can be downloaded from their web site. For cancer biologists and researchers the site offers the following:

- Download genomic and clinical data from a wide variety of cancer types
- Query the database of animal models for human cancers
- Prepare microarray data for analysis
- Analyze proteomics data
- Query and share diverse data types via a web portal
- Organize and design clinical trials^{29, 30}

For more information on translational bioinformatics and related databases in the context of biomedicine, we refer you to the textbook edited by Shortliffe and Cimino.³¹

Future Trends

Two major themes are appearing in the field of translational bioinformatics: (1) the potential for personal genetic ("direct to consumer") services and (2) integration of genomic information into electronic health records.

Personal Genomics

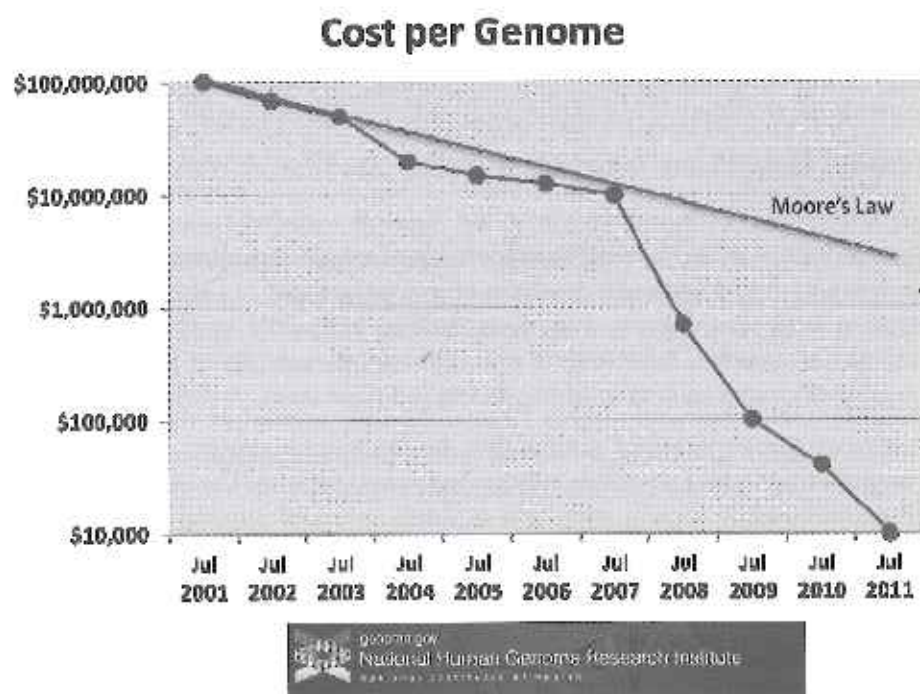
These trends are largely possible because of the availability of population-based genetic data and the decreasing cost for human genome determination.

- **Population Studies:** There are a number of ongoing initiatives that will leverage genomic data in the context of population studies. For instance, Oracle Corporation will partner with the government of Thailand to develop a database to store medical and genetic records. This initiative was undertaken to offer individualized "tailor made" medications and to offer bio-surveillance for future outbreaks of infectious diseases such as avian influenza.³² Not all such initiatives have been successful. Perhaps the best known is DeCODE Genetics Corporation,

which aimed to collect disease, genetic and genealogical data for the entire population of Iceland; however, it filed for chapter 11 bankruptcy in 2009.³³ Nonetheless, DeCODE continues some operations and the development of personal genomics based solutions, largely in partnership with organizations like Pfizer.

- **Decreasing Cost of Human Genome Determination:** Coinciding with the completion of the HGP, the NHGRI has kept track of the cost to perform DNA sequencing of an entire human genome over the past decade. As Figure 22.6 indicates, the cost has dropped from an initial cost of \$100,000,000 to a current cost of about \$10,000 per genome. Notably, the decrease in cost of genome sequence is exceeding Moore's Law (attributed to Intel co-founder Gordon Moore, and states that the cost of computing power will be halved every 18 months based on advances in technology).³⁷

Figure 22.6: Cost per Genome over time (Courtesy National Human Genome Research Institute)



- **Personal Genetics Testing.** Many patients will want to know their own genetic profile, even if the consequences are uncertain. The following are examples of personal genetics companies ("direct to consumer genomics"):
 - Celera Genomics will take advantage of the genomics project to offer genetic mapping services and pharmacogenomics. They offer a cystic fibrosis genotyping assay.³⁴
 - DNA Direct is another company that offers online genetic testing and counseling. They do offer both patient and physician education and have staff genetic counselors.³⁵
 - deCODE Genetics offers whole genomic sequencing as well as deCODEme, an analysis for 47 diseases, traits and ancestry. They also can run risk profiles for type 2 diabetes, prostate cancer, atrial fibrillation, myocardial infarction, glaucoma and breast cancer. A simple mouth wash provides the DNA needed for analysis.³³

- 23andMe is a direct to consumer online genetic testing company. For \$99 they will send a testing kit to homes based on analyzing saliva with a turnaround time of four to six weeks. Currently, they look for 97 diseases, carrier states and drug response conditions. They also offer an analysis of ancestry based on the genetic profile.³⁶ In 2010 a genome wide association study (GWAS) was published that used this technology and showed that patient questionnaire results correlated well with genetic results. Additionally, they were able to describe five new genotype-phenotype associations: freckling, photic sneeze reflex, hair curl and failure to smell asparagus.³⁷ Google's co-founder Sergey Brin has funded a project through this company to study the genetic inheritance of Parkinson's disease. They hope to recruit 10,000 subjects from various organizations and offer a discount price for complete analysis.³⁸

However, as pointed out by Dr. Harold Varmus, personal genetics "is not regulated, lacks external standards for accuracy, has not demonstrated economic viability or clinical benefit and has the potential to mislead customers."³⁹

In order for genetics to enter the mainstream, new technologies and specialties will need to be developed and numerous ethical questions will arise. Just finding the abnormal gene is the starting point. Genetic tests will have to be highly sensitive and specific to be accepted. In general, patients will not be willing to undergo major procedures (e.g., a prophylactic mastectomy or prostatectomy to prevent cancer) unless the genetic testing is nearly perfect. It is also important that genetic counseling be available to help patients understand the implication of genetic susceptibility tests (versus genetic guarantee of disease, such as the mutations associated with Huntington's disease).

Additionally, the Genetic Information Nondiscrimination Act of 2008 was passed to protect patients against discrimination by employers and healthcare insurers based on genetic information. Specifically, the Act prohibits health insurers from denying coverage to a healthy individual or charging that person higher premiums based solely on genetic information and bars employers from using individuals' genetic information when making decisions related to hiring, firing, job placement, or promotion.⁴⁰

Many obstacles face the routine ordering of genetic tests by the average patient. Ioannidis et al. points out that in order for genetic testing to be reasonable several facts must be true. The disease you are interested in must be common. Even with breast cancer, when you evaluate seven established genetic variants, they only explain about 5% of the risk for the cancer. If the disease (e.g., Crohn's disease) is rare, then the test must be highly predictive. In order for genetic testing to be relevant you should have an effective treatment to offer, otherwise there is little benefit. The test must be cost effective, as many currently are too expensive. As an example, screening for sensitivity to the blood thinner warfarin (Coumadin) makes little sense at this time due to cost.⁴¹

A 2010 *Lancet* journal commentary also warned of additional concerns. Whole-genome sequencing will generate a tremendous amount of information that the average physician and patient will not understand without extensive training. At this point, we lack adequate numbers of geneticists and genetic counselors that understand the implications of data being made available thanks to continued advances in biotechnology. Patients will need to sign an informed consent to confirm that many of the findings will have unclear meaning. They will have to deal with the fact that they may be found to be carriers of certain diseases that may have impact on childbearing, etc. Genetic testing may cause many further tests to be ordered, thus leading to increased healthcare expenditures. As we gain more information about whole-genome sequencing, more patients will desire it but who will pay for it? And can the costs be justified?⁴²

Two other recent articles drive home additional practical points. When the risk of cardiovascular disease based on the chromosome 9p21.3 abnormality was evaluated in white women, it only slightly improved the ability to predict cardiovascular disease above standard, well-accepted risk factors.⁴³ Meigs et al. looked at whether multiple genetic abnormalities associated with Type 2 diabetes would be predictive of the disease.

They found that the score based on 18 genetic abnormalities only slightly improved the ability to predict diabetes, compared to commonly accepted risk factors.⁴⁴

For more information regarding future bioinformatics trends we refer to the review paper by Allman and Miller.⁴⁵

Integration with Electronic Health Records

Eventually, the patient's genetic profile will be one more data field in the electronic health record. Recently, gene variants have been identified for diabetes, Crohn's disease, rheumatoid arthritis, bipolar disorder, coronary artery disease and multiple other diseases.⁴⁶ There are a number of forward-looking initiatives that have started on the path to integrate genomic data with traditional clinical data, for example:

- In late 2006 the Veterans Affairs healthcare system began collecting blood to generate genetic data that it will link to its EHR. The goal is to bank 100,000 specimens as a pilot project and link this information to new drug trials. The new voluntary program was officially launched in 2011 and is known as the Million Veteran Program (MVP). MVP will link genetic, military exposure, health and lifestyle into a single database.⁴⁷
- Kaiser Permanente created the Research Program on Genes, Environment and Health and in the first phase two million members will be surveyed to determine their medical history, exercise and eating habits. As of mid-2011 genetic, medical and environmental information had been collected on 100,000 of its members. Kaiser plans correlative studies with its 15 years of digital health information, collected through its electronic health record system. Because the average age of participants is 65 it is anticipated that excellent information about aging will be generated. For example, they are measuring telomere length (the tips of chromosomes) that is thought to correlate with aging. This NIH funded initiative was completed in 15 months, thanks to newer technologies. It is anticipated that data will be analyzed and available to other researchers by 2012.^{48, 49}
- The Electronic Medical Records and Genomics (eMERGE) Network is a consortium of biomedical informatics researchers across the United States. The National Human Genome Research Institute organizes this network, with additional funding from the National Institute of General Medical Sciences. An important theme is whether electronic health records are a vital resource for complex genomic analysis of disease susceptibility and patient outcomes in diverse patient populations.⁵⁰
- Vanderbilt University recently published a strong correlation between their genetic biorepository known as BioVU (genotype) with clinical information (phenotype) obtained from their electronic health record. The diseases studied were rheumatoid arthritis, multiple sclerosis, Crohn's disease and type 2 diabetes.⁵¹

SNOMED CT is making changes to its codes to include genetic information and the National eHealth Initiative is developing "use cases" for family history and genetics so standards can be created by organizations like the Health Information Technology Standards Panel (HITSP). Organizations such as Partners HealthCare, IBM, Cerner and data mining vendors are all gearing up to add genetic information to what we currently know about patients and integrate that with electronic health records.⁵²

The Agency for Healthcare Research and Quality (AHRQ) is developing computer-based clinical decision support tools to help clinicians use genetic information to treat conditions with a strong genetic component, such as breast cancer. Such tools that could be integrated into EHRs are: whether women with a family history of breast cancer need BRCA1/BRCA2 testing and which women who already have breast cancer may benefit from additional genetic testing.⁵³

It is surprising that family history is often overlooked by clinicians and that it usually does not exist as computable data for analysis. To our knowledge, no electronic health record collects this information in a common computable format and uses it for clinical decision support; family history data are generally entered as unstructured text that can be of varying quality (based on provider-patient interviews). Data standards have been developed so family history can be part of EHRs and PHRs, in order to be shared.⁵⁴ There is a government sponsored free web tool available for the public to record their family history using the newest data standards. In this way, the results can be saved as a XML file and shared by EHRs and PHRs. The site, *My Family Health Portrait*, is available for English or Spanish speaking patients, is easy to use and does not store any patient information on the site. Instead, patients can store the XML file on their personal computers.⁵⁵ The program is open source and downloadable from this site.⁵⁶

For further information about the role of EHRs and genomics we refer you to these citations.^{57, 58}

Key Points

- Traditionally, bioinformatics has been a field remote from clinical medicine, but translational bioinformatics will likely bridge this gap
- Advances in biotechnology (such as genome sequencing) will likely introduce a treasure trove of genetic information that will enable deeper understandings of the manifestation of disease as well as the development of a new cadre of therapeutics over the next decade
- The inclusion of genetic profiles is being contemplated for electronic health records
- At this time, direct to consumer genetic testing is still in its early stages, and cannot be used as a replacement for traditional clinical tests (but may be used in complement)

Conclusion

The Human Genome Project and bioinformatics may seem foreign to many clinicians. The promise of translational bioinformatics is to transform biological knowledge (such as can be inferred from genomic data) into clinically actionable items. The success of translational bioinformatics will not be realized until clinicians can access and clinically interpret data that tells them who should be screened for certain conditions and which drugs are effective in which patients as part of day-to-day practice. In the meantime, biomedical scientists and companies will continue to add to the many genetic databases, develop genetic screening tools and get ready for one of the newest revolutions in medicine. The American Health Information Community (AHIC) recommended in 2008 that the federal government should prepare for the storage and integration of genetic information into many facets of healthcare.⁵⁹ Their recommendations will initiate the necessary dialogue that must take place to prepare for bioinformatics to align with the practice of medicine. But, as pointed out by Dr. Varmus "the full potential of a DNA-based transformation of medicine will be realized only gradually, over the course of decades."⁴⁰

References

1. NCBI. A Science Primer. www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html (Accessed July 1 2006)
2. Biotech: Bioinformatics: Introduction www.biotech.icmb.utexas.edu/pages/bioinform/BIintro.html (Accessed July 10 2006)
3. Bioinformatics Overview. Bioinformatics Web www.geocities.com/bioinformaticsweb/?200630/ (Accessed July 6 2006)

Public Health Informatics

ROBERT E. HOYT

JUSTICE MBIZO

NORA J. BAILEY

Learning Objectives

- Define public health informatics
- Define public health surveillance and how data is used in public health
- Explain the significance of information technology in the field of public health
- Explain the significance of syndromic surveillance for early detection of bioterrorism, emerging diseases and other health events
- Explain the significance and scope of global public health informatics
- Understand the workforce needs and competencies of a public health informatician
- List several of the current surveillance systems used in the field of public health
- Explain the function and purpose of the Public Health Information Network

Introduction

Public health is another medical sector that has been greatly influenced by advances in information technology over the past two decades. The overarching goal has been to monitor a variety of medical diseases and conditions rapidly and accurately so as to intervene as early as possible to detect, prevent, and mitigate the spread of epidemics, the effects of natural disasters, and bioterrorism. With the advent of the internet, ubiquitous computing, electronic health records and health information organizations this vision is now possible.

For much of the 20th Century, public health reporting and surveillance consisted of physicians, hospitals and clinics sending paper reports to local health departments, who in turn forwarded information to state health departments who sent the final data to the Centers for Disease Control and Prevention (CDC) via mail or fax and finally to the World Health Organization for certain diseases. Although paper reports are still used, the

shift to electronic media and information technology has facilitated more efficient methods of public health surveillance, community based outbreak detection and disease control.

The most critical component in any disease investigation is the availability of timely data and information to pinpoint the possible source of the outbreak. The proliferation of information technology into public health and medical fields have significantly improved disease surveillance and enhanced early detection of community or population based epidemics. Global events, ranging from the September 11, 2001 terrorist attacks, the emergence of severe acute respiratory syndrome (SARS) in 2002 in China, to the recent global H1N1 influenza outbreak reinforced the need for robust interoperable surveillance systems. The terrorist events of September 11, 2001 in particular, the subsequent anthrax attacks across the United States elevated and reinforced public health to a national security issue increasing the need for biosurveillance and real-time data analysis to detect and respond to disease outbreaks and health events more rapidly.

In the following sections we will define public health informatics (PHI), discuss public health surveillance systems, discuss syndromic surveillance, geographic information systems and cover global public health informatics.

Definitions

- Public health: the science and art of preventing disease, prolonging life and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals."¹
- Public health informatics: "the systematic application of information and computer science and technology to public health practice, research and learning...."²
- Public health surveillance: "the ongoing systematic collection, analysis, and interpretation of health-related data essential to the planning, implementation and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know. The final link in the surveillance chain is the application of these data to prevention and control."³
- Syndromic surveillance: "surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response."⁴

Public Health Surveillance

Public health surveillance is essential to understanding the health of a population. Until recent years, public health surveillance was primarily paper-based. However, with the increasing shift towards eHealth public health surveillance has embraced the field of public health informatics.¹² In order to study a large population we need interoperable technologies such as standards-based networks, databases and reporting software. Current electronic surveillance systems employ complex information technology and embedded statistical methods to gather and process large amounts of data and to display the information for networks of individuals and organizations at all levels of public health. Public health surveillance serves to:

- Estimate the significance of the problem
- Determine the distribution of illness
- Outline the natural history of a disease
- Detect epidemics
- Identify epidemiological and laboratory research needs

- Evaluate programs and control measures
- Detect changes in infectious diseases
- Monitor changes in health practices and behaviors
- Assess the quality and safety of health care, drugs, devices, diagnostics and procedures
- Support planning¹³

Types of Surveillance Systems

Public health surveillance systems can be classified based on data collection purpose and design. Table 23.1 demonstrates the more common categories.¹⁴⁻¹⁸

Table 23.1: Types of Surveillance Systems

Surveillance System	Definition/Description	Examples
Case surveillance systems	<ul style="list-style-type: none"> • Collect data on individual cases of a health event or disease with previously determined case definitions in respect to criteria for person, time, place, clinical & laboratory diagnosis • Analyze case counts and rates, trends over time and geographic clustering patterns • Historically, case surveillance has been the focus of most public health surveillance. 	<ul style="list-style-type: none"> • National Notifiable Disease Surveillance System (NNSS)
Syndromic surveillance systems	<ul style="list-style-type: none"> • Collect data on clusters of symptoms and clinical features of an undiagnosed disease or health event in near real time allowing for early detection, rapid response mobilization and reduced morbidity and mortality • Data can be obtained through specific surveillance systems as well as existing epidemiologic data such as insurance claims, school and work absenteeism reports, over the counter (OTC) medication sales, consumer driven health inquiries on the Internet, mortality reports and animal illnesses or deaths for syndromic surveillance. • Geographic and temporal aberration and geographic clustering analyses are performed with real-time syndromic surveillance data. • Syndromic surveillance systems can also be used to track longitudinal data and monitor disease trends. 	<ul style="list-style-type: none"> • Real-time Outbreak Detection System (RODS) • Biosurveillance Common Operating Network (BCON) • BioSense 2.0
Sentinel surveillance systems	<ul style="list-style-type: none"> • Collect and analyze data from designated agencies selected for their geographic location, medical specialty, and ability to accurately diagnose and report high quality data. They include health facilities or laboratories in selected locations that report all cases of a certain health event or disease to analyze trends in the entire population. • Pros: Useful to monitor and identify suspected health events or diseases • Cons: Less reliable in assessing the magnitude of health events on a national level as well as rare events since data collection is limited to specific geographic locations. 	<ul style="list-style-type: none"> • PulseNet • FoodNet • ILINet
Behavioral surveillance systems	<ul style="list-style-type: none"> • Collect data on health-risk behaviors, preventative health behaviors, and health care access in relation to chronic disease and injury. • Analyze the prevalence of behaviors as well as the trends in the prevalence of behaviors over time. • Information is most commonly collected by personal interview or examination • Inferential and descriptive analysis methods such as age-adjusted rates, linear regression, and weighted analyses are used. • Most acute when conducted regularly, every 3 to 5 years 	<ul style="list-style-type: none"> • Behavioral Risk Factor Surveillance System (BRFSS) • Youth Risk Behavior Surveillance System (YRBSS) • National Health Interview Survey (NHIS) • Pregnancy Risk Assessment Monitoring System (PRAMS)

Surveillance System	Definition/Description	Examples
Integrated Disease Surveillance and Response (IDSR)	<ul style="list-style-type: none"> Incorporates epidemiologic and laboratory data in systems designed to monitor communicable diseases at all levels of the public health jurisdiction, particularly in Africa. Useful for: detecting, registering and confirming individual cases of disease; reporting, analysis, use, and feedback of data; and preparing for and responding to epidemics. 	
Clinical Outcomes Surveillance	<ul style="list-style-type: none"> Monitors clinical outcomes to study disease progression or regression in a population. Analyzes the rates of and factors associated with clinical outcomes using descriptive and inferential methods such as incidence rates from probability samples. 	<ul style="list-style-type: none"> Medical Monitoring Project that monitors and tracks HIV patients
Laboratory Based Surveillance	<ul style="list-style-type: none"> Collects data from public health laboratories, which routinely conduct tests for viruses, bacteria, and other pathogens. Used to detect and monitor infectious and food-borne diseases based on standard methods for identifying and reporting the genetic makeup of specific disease-causing agents. Commonly used in case surveillance and sentinel surveillance. 	<ul style="list-style-type: none"> PulseNet National Case Surveillance for Enteric Bacterial Disease (CDC)

The CDC has a helpful web page dedicated to surveillance programs for state, tribal, local and territorial public health officials.¹⁹

Syndromic Surveillance

Syndromic surveillance is part of meaningful use; therefore a basic understanding is important. Syndromic surveillance means symptoms are monitored (like diarrhea or cough) before an actual diagnosis is made. If, for example, multiple individuals complain of stomach symptoms over a short period of time, one can assume there is an outbreak of gastroenteritis. The important thing to remember is that syndromic surveillance systems do not identify the cause of the outbreak, rather they provide data comparisons which allows public health official to initiate outbreak investigation techniques.

In addition to the obvious sources of health data, public health officials can also monitor and analyze: unexplained deaths, insurance claims, school absenteeism, work absenteeism, over the counter medication sales, Internet based health inquiries by the public and animal illnesses or deaths.¹⁵

Initially, public health officials were very interested in detecting trends or epidemics in infectious diseases, such as severe acute respiratory syndrome (SARS) and avian influenza. After the terrorist attacks and anthrax outbreak in 2001, they have had to improve biosurveillance to detect bioterrorism. The objective is to "identify illness clusters early, before diagnoses are confirmed and reported to public health agencies and to mobilize a rapid response, thereby reducing morbidity and mortality."²⁰ The challenge is to develop elaborate systems that can sort through the information and reduce the signal to noise ratio. The syndrome categories most commonly monitored are:

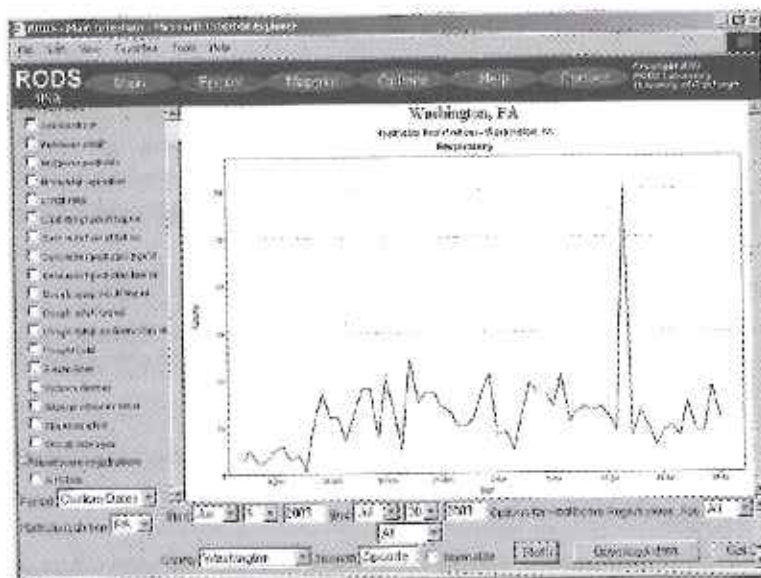
- Botulism-like illnesses
- Febrile (fever) illnesses (influenza-like illnesses)
- Gastrointestinal (stomach) symptoms
- Hemorrhagic (bleeding) illnesses
- Neurological syndromes
- Rash associated illnesses
- Respiratory syndromes
- Shock or coma

Ambulatory electronic health records (EHRs) are a potentially rich source of data that can be used to track disease trends and biosurveillance. EHRs contain both structured (e.g. ICD-9 coded) data as well as narrative free text. Hripesak et al. assessed the value of outpatient EHR data for syndromic surveillance. Specifically, they developed systems to identify influenza-like illnesses and gastrointestinal infectious illnesses from Epic® EHR data from 13 community health centers. The first system analyzed structured EHR data and the second used natural language processing (MedLEE processor) of narrative data. The two systems were compared to influenza lab isolates and to a verified emergency room (ER) department surveillance system based on "chief complaint." The results showed that for influenza-like illnesses the structured and narrative data correlated well with proven cases of influenza and ER data. For gastrointestinal infectious diseases, the structured data correlated very well but the narrative data correlated less well. They concluded that EHR structured data was a reasonable source of biosurveillance data.²¹

Real-Time Outbreaks Detection System (RODS)

The RODS system was initially developed by researchers at the University of Pittsburgh and was the first real-time detection system for outbreaks. RODS collected patient chief complaint data from eight hospitals in a single health-care system via Health Level 7 (HL7) messages in real time, categorized these data into syndrome categories by using a classifier based on International Classification of Diseases, Ninth Revision (ICD-9) codes, aggregated the data into daily syndrome counts and analyzed the data for anomalies possibly indicative of disease outbreaks. Much like the ESSENCE system, RODS system started with a set of mutually exclusive and exhaustive categories of eight syndromic categories. However, as the program has gone through revisions and refinement, the categories have been reduced to seven as follows: respiratory, gastrointestinal, botulinic, constitutional, neurologic, rash and hemorrhagic. Figure 23.1 shows the daily counts of respiratory cases for Washington County, PA in the period June-July 2003.

Figure 23.1: Daily counts of respiratory cases six month period, Washington County, PA 2003.



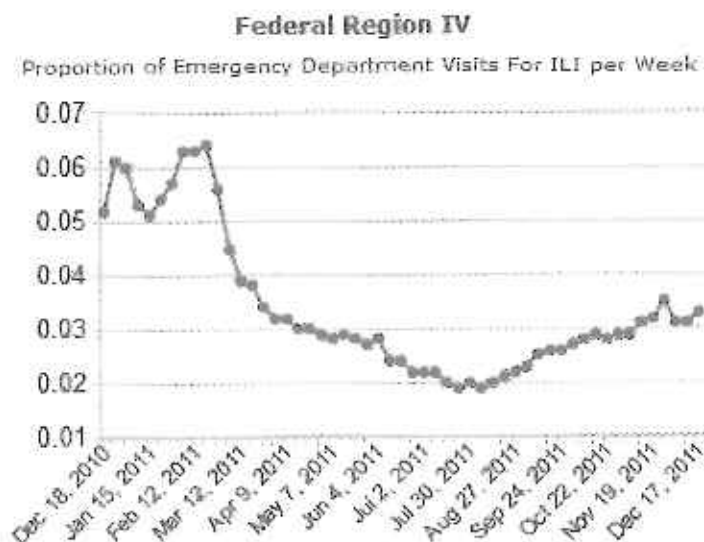
Sources: Real-Time Outbreak and Disease Surveillance project.
 * The June 2003 increase corresponds to new hospitals being added to the system.
 † The sudden increase on July 18, 2003, was caused by 60 persons reporting to one emergency department within 4 hours for carbon monoxide exposure.

In order to increase the adoption of the RODS system, the University of Pittsburgh started offering software free of charge to public health departments. In 2003 the software was offered under an open source license and since then many more agencies have adopted the software for their use.²²

Distribute

This project was created by the International Society for Disease Surveillance (www.syndromic.org), with the goal of supporting emergency department (ED) surveillance of influenza like illnesses (ILI). Figure 23.2 shows ILI reported over the last year in south eastern United States (region IV).²³

Figure 23.2 Proportion of ED visits for ILI weekly 2011 (Courtesy Distribute)



BioSense

This is a CDC national web-based program to improve disease detection, monitoring and situational awareness for healthcare organizations in the United States by reporting emergency room data. Participants include DOD, VA and civilian hospitals. The program addresses identification, tracking and management of naturally occurring events as well as bioterrorism. In 2010 BioSense was redesigned to integrate existing syndromic surveillance systems and allow for better regional sharing of information. The 2011 BioSense 2.0 allows state and local health departments to access data that will support syndromic surveillance systems under meaningful use. The goal is to provide a web based clearinghouse where data can be stored, searched and analyzed from and by multiple parties; decreasing the need for local health departments to purchase additional expensive information technologies.²⁴

The Public Health Information Network

The Prevention and Public Health Fund, as part of the Affordable Healthcare Act of 2010, in conjunction with the Health Information Technology for Economic and Clinical Health (HITECH) Act has allowed the public health infrastructure to move into the eHealth era. Driven by the mission to prevent, reduce and treat disease, these initiatives focus on developing interoperable public health information systems that are beneficial to the healthcare of all Americans.⁵⁻⁶

The Public Health Information Network (PHIN) is a Centers for Disease Control and Prevention (CDC) initiative established to provide the framework for efficient public health information access, exchange, use, and collaboration among multi-level public health agencies and partners using a consensus of shared policies, standards, best practices, and services.⁷

Establishing messaging and vocabulary standards is a key strategy for PHIN, allowing for consistent interoperability between local, state and national public health entities as well as other agencies. The PHIN is currently working with the following Standard Development Organizations (SDOs): Systematic

Nomenclature for Medicine (SNOMED), Logical Observation Identifiers Names and Codes (LOINC), Health Level 7 (HL7), and Consolidated Health Informatics Initiative (CHII).⁸ For more information about data standards, we refer readers to Chapter 6.

Electronic Health Records

Integral to this vision is interoperability with electronic health records (EHRs), as part of Meaningful Use of Health IT. As stated in the chapter on EHRs, Meaningful Use Stage 1 has several menu objectives with public health implications: the capability to transmit syndromic surveillance data to public health agencies, the capability to transmit data to immunization registries, and the capability for hospitals to transmit required disease Electronic Laboratory Reports (ELRs).¹⁰

As we see increased adoption of EHRs in the US and progression to Meaningful Use Stages 2 and 3, we should start to see new sources of data available for public health analysis.¹⁰

Public Health Information Network Update

On October 18, 2011, the "PHIN Messaging Guide for Syndromic Surveillance: Emergency Department and Urgent Care Data Version 1.0" was released, reaching a milestone for the public health objectives of Meaningful Use under the HITECH Act. This guide provides the HL7 2.5.1 messaging standard for the use of emergency department data as syndromic surveillance.⁹

Health Information Exchange (HIE)

We anticipate more public health reporting as a result of Meaningful Use for EHRs but a broader approach would be aggregating EHR/data shared with a health information organization. For further information about health information exchange we refer readers to Chapter 5.

A recent article outlined use cases that demonstrate the utility of HIE in public health:

- **Mandated reporting of lab diagnoses:** there is a predefined list of *notifiable diseases* (e.g. TB) that would benefit from electronic transmission to public health. In spite of that many states still rely on paper and results must be mapped to a standard vocabulary such as LOINC. A health information organization (HIO) could ensure proper identification, archiving and mapping. Mandated reporting could also trigger an alert of reportable diseases.
- **Non-mandated reporting of lab data:** There are several infectious diseases of interest that are not on the *notifiable* list but ideally tracked by public health. Additionally, antibiotic resistance patterns should be reported and shared with public health. A community wide antibiogram could be developed to educate local physicians about optimal prescribing patterns.
- **Mandated reporting of physician-based diagnoses:** physicians are separately required to report certain *notifiable diseases* but reporting is highly variable. This could be made easier with EHR reporting to the local HIO that in turn reports to public health. Data standards would be essential and alerts to appropriate public health staff, infection control officers, etc. would be possible.
- **Non-mandated reporting of clinical data:** syndromic surveillance will require symptom-related data from EHRs and emergency departments (EDs) to be sent and analyzed.
- **Public health investigation:** public health officials could query the HIO for additional clinical or demographic (age, gender, location, etc.) information about a case of interest.

- **Clinical care in public health clinics:** clinicians who treat patients in public health clinics could potentially benefit from access to a HIO.
- **Population-level quality monitoring:** HIE has the potential to give public health officials a glimpse of the quality of medical care in their area without chart reviews, across multiple health care systems.
- **Mass-casualty events:** HIOs might serve as a single point of contact for victims of a mass casualty. A record locator service might be able to keep track of admissions, discharges and transfer (ADT) data for the victims and their families.
- **Disaster medical response:** HIOs have the potential to make available patient data during a disaster when paper records might be destroyed or unavailable.
- **Public health alerting—patient level:** Theoretically, public health departments could alert all clinicians in a HIO about a case of TB where follow up is lost, for example. Public health officials could also warn hospitals about unique cases of highly resistant infectious organisms, particularly when patients tend to seek medical care at multiple institutions.
- **Public health alerting - population level:** Clinicians could be warned about trends in the community, for example viral culture results or antibiotic resistance trends.¹¹

Geographic Information Systems (GIS)

As early as 1855 Dr. John Snow created a simple map to show where patients with cholera lived in London in relation to the drinking water source in the Soho District of London. Using his hand drawn map and basic epidemiological investigation techniques, much of which are still used today, he determined the source of the epidemic to be a common water pump. Epidemiology, public health surveillance and indeed the field of public health have improved significantly since the pioneer work of Snow and others after him. Much of this transformation has been the result of the emergence and proliferation of advanced computing technologies, the internet and other automated information systems.

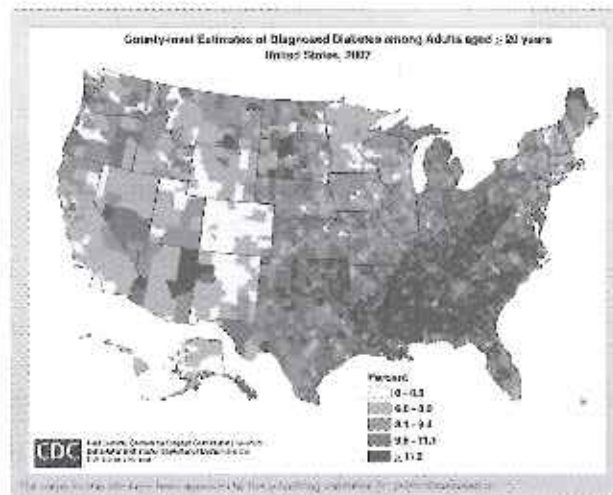
Modern geographic information systems (GIS) use digitized maps from satellites or aerial photography. A Geographic Information System (GIS) is a system of hardware, software and data used for the mapping and analysis of geographic data. GIS provides access to large volumes of data; the ability to select, query, merge and spatially analyze data; and visually display data through maps. GIS can also provide geographic locations, trends, conditions and spatial patterns. Spatial data has a specific location such as longitude-latitude, whereas attribute data is the database that describes a feature on the map.

GIS maps are created by adding layers. Each layer on a GIS map has an attribute table that describes the layer. The data can be of two types: *Vector* or *Raster*. *Vector* data appears as points, lines or polygons (enclosed areas that have a perimeter like parcels of land). *Raster* data utilizes aerial photography and satellite imagery as a layer. Using GPS and mobile technology, field workers can enter epidemiologic data to populate a GIS. This geospatial visualization has been useful in tracking infectious diseases, public health disasters and bioterrorism.^{15, 26}

With the recent shift in public health focus to preventable chronic diseases, GIS has also been used to monitor chronic diseases and social and environmental determinants of health for public health policy. In early 2011, the Centers for Disease Control and Prevention launched a new project, Chronic Disease GIS Exchange. Designed for public health professionals and community leaders, GIS experts will use as an information exchange forum to network and collaborate with the goal of preventing heart disease, stroke and other chronic diseases. Data and information shared in this forum will be used in documenting the disease

burden, informing policy decisions, enhancing partnerships and facilitating interventions from the use of GIS data.²⁷⁻²⁹ Figure 23.3 shows a GIS display of diabetes incidence rates by State.²⁸

Figure 23.3: GIS Map of diabetes diagnosis by county (Courtesy CDC Chronic Disease GIS Exchange)



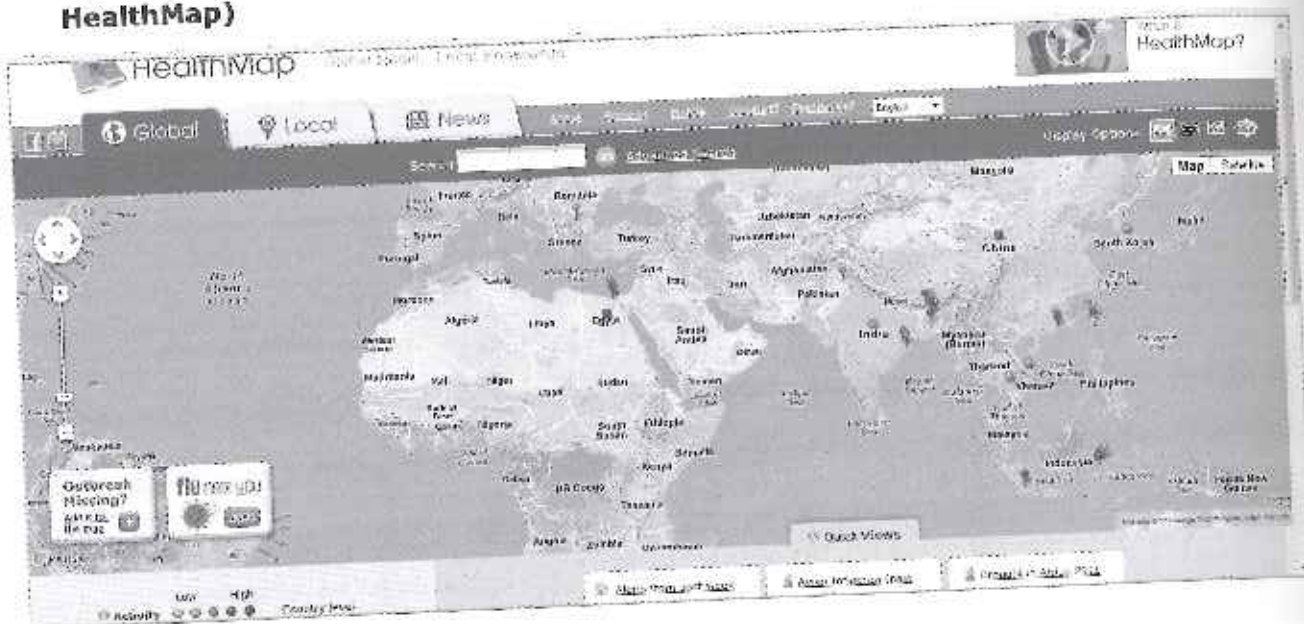
Virtually all of the biodetection systems mentioned have a GIS component that allows for the mapping of disease outbreak events giving public health practitioners the ability to timely deploy resources to control the outbreak and prevent further spread. Key variables can be inputted by zip code, latitude, longitude, that help public health disease investigators narrow down the source of the problem.

IhealthMap is a global project to integrate infectious disease news and visualization using an Internet geographic map. This program classifies alerts by location and disease. For example, you can select “malaria” and “global” and see if there were any reported cases in the past 30 days. “Mouseover” an icon and you will see what is being reported in that area. A smartphone app “Outbreaks near me” details H1N1 (swine flu) outbreaks by locale, in near real time.³⁰ The program was developed by the Harvard-MIT Division of Health Sciences and Technology and a more detailed explanation of the system and architecture is provided at this reference.³¹ Figure 23.4 shows a GIS display of global avian flu outbreaks

Public Health Informatics Workforce

As discussed, in order to most accurately and efficiently study the health of the population, information and communication technologies are essential to support the increasing demand for public health research and evidence based public health practice as a result of the aging US population. These technologies also require a diversity of human expertise for management, analysis, and communication of public health data. The Association of Schools of Public Health (ASPH) estimates that the field of public health will require 250,000 more workers by 2020 to avert a national public health crisis.³² The transition to eHealth requires all public health workers to have some knowledge of IT depending on the demands of their position. In anticipation of this need, the CDC in collaboration with the University of Washington’s School of Public Health and Community Medicine’s Center for Public Health Informatics developed a list of informatics competencies for public health workers to meet the needs of the evolving public health field as well as for the Public Health Informatician. A Public Health Informatician is “a public health professional who works in practice, research, or academia and whose primary work function is to use informatics to improve population health.”³³

Figure 23.4: GIS display of global avian influenza outbreaks (Courtesy HealthMap)



Global Public Health Informatics

Public health threats from chronic and infectious diseases, population health status, and health disparities within and across countries have gained global attention in part due to increasing personal mobility, economic globalization, and expansion of communication technologies. In fact, the global threat from chronic diseases was the focus of the 2011 UN General Assembly. Infectious diseases, such as influenza, polio, and HIV/AIDS, can quickly spread across national borders and are best curtailed through international cooperation and timely information sharing. New or re-purposed health information technologies provide critical support in the identification, monitoring, alerting, and responding to emerging diseases, pandemics, bioterrorism, and natural disasters. Simultaneously, health informatics has also emerged as an important tool in addressing population health goals and as a means to reduce health disparities between *developed* and *developing* nations.

World Health Organization

The leading international public health entity is the World Health Organization (WHO). Organized in 1948 as an agency of the United Nations (UN), WHO directs and coordinates public health efforts worldwide. WHO and its 195 Member States collaborate with other UN agencies, nongovernmental organizations, and the private sector to:

- **Foster health security:** Through its surveillance and disaster/epidemic response systems, WHO works to identify and curb outbreaks of emerging or epidemic-prone diseases. The revised 2007 International Health Regulations address the major forces contributing to epidemics including urbanization, environmental mismanagement, food preparation, and the overuse of antibiotics.
- **Promote health development:** Through this objective WHO works to increase access to life-saving and health-promoting interventions, particularly in poor, disadvantaged, or vulnerable groups. WHO's health development efforts focus on the treatment of chronic and infectious disease (e.g. diabetes), prevention and treatment of tropical diseases (e.g. malaria), women's health issues, and healthcare within African nations.

- **Strengthen health systems:** In poor and medically underserved areas, WHO endeavors to strengthen and supplement existing health systems. Activities include providing trained healthcare workers, access to essential drugs, and assistance in collecting vital health information.³⁴

As discussed throughout this section, WHO increasingly relies on health information technology to carry out its objectives.

International Surveillance and Response Programs

The most visible role of WHO is to detect and respond to infectious disease outbreaks, pandemics, and disaster emergencies. Global surveillance of infectious disease, famines, and environmental disasters is implemented through a network of regional, national, and international institutes.³² Government organizations (e.g. CDC), military networks (e.g. US Department of Defense's Global Emerging Infections Surveillance and Response System), and a host of public and private non-governmental organizations (NGOs) (e.g. Google, HealthMaps) monitor and report infectious diseases to WHO. Additionally, internet sites such as Epi-X or Pro-Med maintain discussions on current infectious diseases.

A 2007 review of 15 international surveillance and response programs (ISRPs) classified their activities into four basic components: surveillance, reporting, verification, and response.³⁵ The report found that the majority of these ISRPs focus on surveillance and reporting, while only six carry out all four activities. These six ISRP as well as other leading surveillance systems are described in the Appendix.

Regardless of the surveillance component performed by an ISRP, these organizations have benefited from the expansion of health information technology into the surveillance arena. Over the past decade, WHO and ISRPs have embraced web-based computing, mobile applications, GIS, and even text messaging. The role of health informatics within the major global surveillance organizations are discussed below.

Global Alert and Response (GAR): GAR is the integrated infectious disease surveillance program within WHO. A network of national, regional, and international agencies, governmental organizations (e.g. CDC) and military networks (e.g. US Department of Defense's Global Emerging Infectious Disease), GAR's primary function is the facilitation of epidemic preparedness and response worldwide. This body is also responsible for maintaining and enhancing the global outbreak and bio-risk operational platforms.³⁶ Global monitoring and coordination are increasingly important in light of recent public health challenges such as outbreaks of severe acute respiratory syndrome (SARS) and influenza A (H1N1), the AIDS epidemic, and emerging new diseases and pathogens. Electronic surveillance capabilities have greatly enhanced the ability of GAR and its component functions to identify and respond to public health emergencies. Subsidiary functions under GAR include:

- **International Health Regulations:** 2005 revisions to WHO's International Health Regulations (IHR) are aimed at improving global public health security and collaborative response to natural disasters, biological or chemical agents, and radioactive material release.³⁷ This legally-binding agreement provides a framework for the management of international public health emergencies, while also addressing the capacity of participating nations to detect, evaluate, alert, and respond to public health events. IHR specifies operational procedures for disease surveillance, notification and reporting of public health events and risks as well as for the coordination of international response to those events. The 2005 IHR allowed for the first time non-governmental sources to provide surveillance information to WHO. Participation by non-governmental contributors is as a positive step that pushes WHO to become more "dynamic, flexible, and forward-looking."³⁸
- **Early Warning Surveillance:** GAR implemented an early warning surveillance response (EWARN) mechanism to effectively identify disease outbreaks and other health issues immediately following acute emergencies. An initial version of the system has been in use in

Haiti since the 2008 hurricane and expanded following the devastating Haitian earthquake in 2010. The system monitored public health issues such as injuries, mental health concerns, TB and HIV treatment programs, and disease trends. Inconsistency of data reporting, lack of trained personnel for data collection and technological errors among other problems interfered with the project from the start.³⁹ One solution that was developed in response to these challenges was a "virtual Google group" set up to improve communication. In remote, undeveloped areas of the world, WHO has encouraged Member States to develop early warning systems that use a variety of media including fax, telephone, the internet, and SMS to connect district or national surveillance officers with field collection efforts.

Global Public Health Intelligence Network (GPHIN): GPHIN was developed by the Public Health Agency of Canada to electronically monitor infectious disease outbreaks. Approximately 40 percent of the outbreaks investigated by WHO each year come from the GPHIN. This network "is a secure, internet-based 'early warning' system that gathers preliminary reports of public health significance in seven languages on a real-time, 24/7 basis."⁴⁰ GPHIN "continuously and systematically crawls web sites, news wires, local online newspapers, public health email services and electronic discussion groups for key words."⁴¹ Although originally developed to detect infectious disease outbreaks, GPHIN now scans for food and water contamination, exposure to chemical and radioactive agents, bioterrorism, and natural disasters. It uses automated analysis to process the gathered data to alert human analysts to conduct additional review of any serious issues or trends. These data are then made available to WHO/GOARN and other subscribers through its web-based Microsoft/Java application and to the public through the WHO web site. GPHIN's automated data has significantly accelerated global outbreak detection.

Malaysia: Early Warning And Risk Navigation Systems



eWARNS is Malaysia's Early Warning And Risk Navigation Systems for natural disasters including rainfall, flash flood, soil erosion, landslide, tidal wave, and forest fire. Remote Sensing and Transmission Units (RSTU) placed throughout the country are used to predict floods and other natural disasters. Each RSTU collects rainfall data, senses the impact of the rain, and transmits the data via the internet to a receiving unit. The RSTU also acts as a web-server allowing the 'remote panel' to be viewed via the internet. The system alerts the public to real time risk levels and forecasts via SMS text messaging on their mobile phones.¹⁰ Information on daily rainfall, erosivity index, and erosion hazards are also available on the website.

<http://www.cwarns.com.my/index.php?im=about>

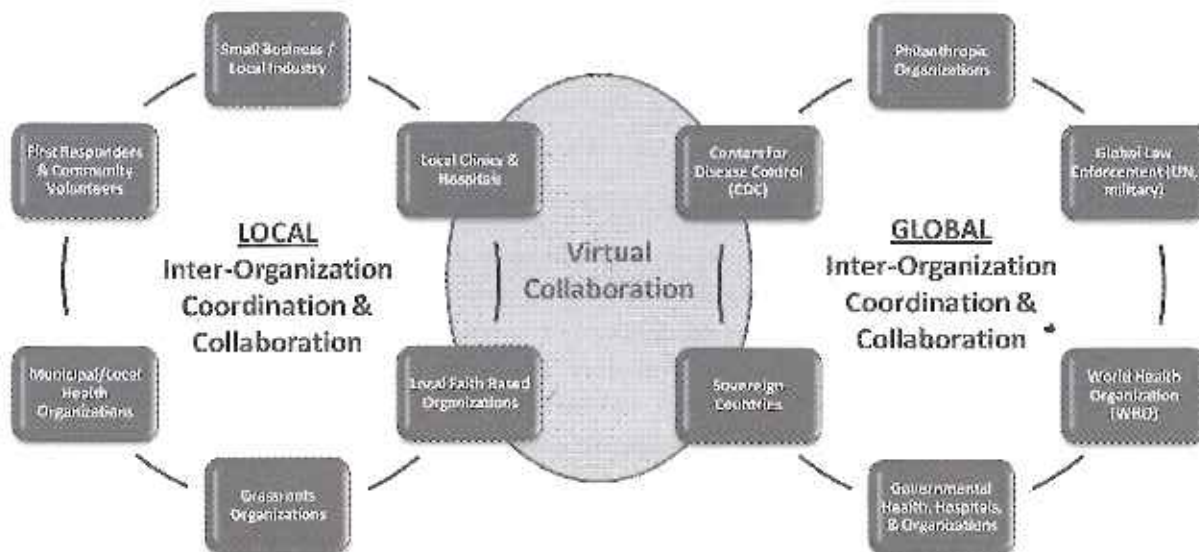
Global Outbreak Alert and Response Network (GOARN): The Global Outbreak Alert and Response Network was established by WHO in 1997. GOARN has 420 global partners to collaboratively provide a rapid identification and response to outbreaks and alert the international community. Collaboration is provided by organizations like the Red Cross, the United Nations, humanitarian and scientific institutions, technical networks, laboratories, and surveillance and medical initiatives.¹²

Since 2000 GOARN has responded to more than 50 events worldwide, including SARS, Avian influenza and H1N1 influenza outbreak. Over one third of the surveillance information coming into GOARN is provided by GPHIN. Other surveillance information is provided by governmental agencies, universities, military agencies, and non-governmental organizations (NGO), such as the Red Cross and Médecins sans Frontières (Doctors without Borders). To facilitate global coordination, GOARN has established standardized operating procedures to be used by Member States and partnering organizations for identifying and responding to outbreaks. Features of the system include: alerts to the international community about outbreaks and

technical collaboration on the rapid identification and response to outbreaks.⁴² WHO's state of the art IT and communications systems ensure secure timely communications within GOARN and between GOARN and Member States and partnering entities thus facilitating the quick response and control of disease outbreaks.

Effective communication and collaboration between local and global responders to public health crises, hazards, and pandemics, is critical to successfully address the complex and diverse needs of the population after a disaster or during a public health emergency. GOARN and other responders recognize the benefit of integrating information and communication technology (ICT) into current operations.⁴³ Although sharing protocols of ICT appear to be a challenge, the emerging field of community informatics seems to provide the potential for inclusion of local health providers in emergency response efforts coordinated by global public health agencies.⁴⁴ Figure 23.5 depicts the complex and interdependent communication that must occur to ensure coordination of the local and global public health entities involved in disaster or public health emergency response.

Figure 23.5: Coordination between local and global public health organizations



Other Global Public Health Activities

Surveillance and response to emergent health events maybe the most visible, but they are not the only functions of public health organizations. Public health is responsible for the prevention and control of disease, chronic and communicable diseases such as HIV/AIDS, TB, and polio and also plays a key role in health promotion and education. Historically, WHO and other public health organizations have struggled to provide even the most basic services to remote and poor areas around the globe. Health technology, particularly *mhealth*, has enabled public health agencies to reach out to isolated villages, connect with paraprofessional field workers, collect data, diagnosis disease, deliver disease management instructions, provide proficiency training to healthcare workers, and educate patients. Some of the organizations that deploy health technology in the fight to improve global health are identified in Table 23.2 on following page.

Global Health Information Technology Programs

Listed in alphabetical order below are a few of the premier organizations that facilitate the use of health information technology for public health:

- Center for Innovation in Global Health Technologies (CIGIT): A component of the Robert R. McCormick School of Engineering and Applied Science at Northwestern University, CIGIT collaborates with other universities, global healthcare companies, and non-profit organizations on the research and development of innovative and affordable healthcare technologies. The

program focuses on three areas that are of concern in developing nations: HIV and associated diseases, saving lives at birth, and training healthcare workers to supplement physicians and nurses. <http://www.cmc.northwestern.edu/global-health-initiatives/index.html>

- **FIMI360-SATELLIFE:** Created in 1987, SATELLIFE is a leader in using information technology to connect healthcare providers in developing nations to vital medical knowledge. Its GATHERdata™ project uses mobile devices to collect, report, and analyze real-time disease surveillance data. <http://www.healthnet.org/>
- **Global Public Health Informatics Program (GPHIP):** The Centers for Disease and Control (CDC) established a Global Public Health Informatics Program (GPHIP) in 2008 to collaborate with WHO and other international partners. "The Goal of GPHIP is to improve domestic and international public health informatics programs and advance the best informatics science, principles, strategies, standards, and practices."⁴⁶ GPHIP assists CDC-supported countries on developing and implementing innovative public health informatics solutions. Collaborative projects supported by GPHIP include a mobile-based information system for use in health emergencies and for surveys in China, an electronic integrated disease surveillance systems (EIDSS) in cooperation with Armenia, Azerbaijan, Georgia, Kazakhstan, Saudi Arabia, Ukraine, and Uzbekistan, and a national disease surveillance (NDS) and a health surveillance network (HISN) in Saudi Arabia. <http://www.cdc.gov/globalhealth/programs/informatics.htm>
- **Information and Communication Technologies for Public Health Emergency Management (ICT4PHEM):** Established by GAVI in 2009, ICT4PHEM "is a technical collaboration of existing institutions and networks who pool human, technical and technological resources together to provide enhanced ICT solutions to predict, prevent and support Public Health Emergencies."⁴⁷ The objective of ICT4PHEM is to deploy ICT in the detection, assessment, verification and response to public health threats throughout the world. The initial meeting was held in April 2009 to discuss the need to develop, enhance and make available ICT tools to public health entities worldwide. <http://www.who.int/csr/ict4phem/en/index.html>
- **WHO Global Observatory for eHealth (GOe):** In 2005, the 58th World Health Assembly recognizing the need to incorporate emerging health information technologies into WHO and Member States adopted an eHealth strategy resolution. That same year WHO established GOe to study the impact of ehealth. The GOe conducted a survey of members in 2005 to establish a benchmark for each nation on its ehealth; a follow-up survey was conducted in 2009. Information on their findings relative to mobile technology, telemedicine, safety and security and other ehealth issues are available on their website: <http://www.who.int/entity/goe/>.
- **Wireless Reach™:** Through its Wireless Reach™ program, Qualcomm works with global partners to bring wireless technology to poor and remote areas around the world. Wireless Reach™ addresses education, entrepreneurship, public safety, and environment in addition to health. Its projects tend to be telemedicine related, although some have public health applicability. <http://www.qualcomm.com/citizenship/wireless-reach>

Table 23.2: Global Efforts to Improve Public Health through the use of Health Information Technology

Organization	Public Health Informatics Services
Cell-Life http://www.cell-life.org/	A not-for-profit organization that deploys mobile technology in the fight against HIV and other communicable diseases, primarily in South Africa. It has effectively used SMS to encourage HIV testing, to remind women to continue in prevention programs to curb mother-to-child transmission of HIV, increase antiretroviral therapy adherence, and provide family planning information.
Datadyne http://www.datadyne.org/	Datadyne offers applications for the use of cell phones to collect data, sending of mass SMS messages, and to provide continuing education to healthcare workers in remote areas through mobile devices.
Dimagi http://www.dimagi.com/	Dimagi is a for-profit company that builds custom mobile health and SMS solutions for resource-poor environments. It offers Windows Mobile 5 software devices to assist community health workers to screen HIV/AIDS patients, personalized SMS medication reminders to increase antiretroviral adherence in HIV patients, a mobile solution to improve home-based cancer care coordination, a portable web application for remote clinics to send cancer screening images to hospital-based physicians, SMS alerts for critical events, mobile applications for continuing education of remote healthcare workers, and a mobile application to increase compliance with WHO's Integrated Management of Childhood Illness program by remote health workers.
E Health Point http://ehealthpoint.com/?page_id=77	This project uses telemedicine to connect rural Indian villages to physicians and evidence based healthcare.
Mobile Alliance for Maternal Action (MAMA) http://www.mobilemamaalliance.org/	MAMA is a public-private partnership involving the US Agency for International Development, Johnson & Johnson, the United Nations Foundation, mHealth Alliance and BabyCenter. MAMA uses mobile phones to send audio and text health messages and reminders to new and expectant mothers.
mHealth Alliance http://www.mhealthalliance.org/	The mHealth Alliance is a public-private partnership between the UN Foundation, the Rockefeller Foundation, and The Vodafone Foundation. Its purpose is to harness the power of wireless technologies to improve health outcomes in low and middle income countries.
WHITIA-Essential Technologies for Safety Net Providers http://www.worldhealthimaging.org/index.html	WHITIA developed a low-cost, simple to use, self-contained digital x-ray unit. These units were initially deployed in Guatemala. WHITIA also provides telemedicine technology to connect village medical personnel with specialists and technology to enable high speed transmission of teleradiology images and healthcare data.

Future Trends

At the core of public health informatics is surveillance, a practice that relies on near-real time, high quality data. Largely because of the increased global use of the Internet, we are seeing an increase in analysis of aggregated data collected by both public and private organizations such as Google and various social media sites like Twitter and Facebook. Google.org recently launched three Internet-based projects utilizing

revolutionary technology for public health research and policy development: *Google Flu Trends*, *Google Dengue Trends*, and *Google Crisis Response*. *Google Flu Trends* and *Dengue Trends* use aggregated data based on Google search queries to estimate disease activity in real-time.⁴⁸⁻⁴⁹ Correlating strongly with data from the CDC, *Google Flu Trends* data is estimated to precede CDC results by about one week.⁵⁰⁻⁵¹ Ultimately, this methodology may be shown to be the most effective and fastest way to identify pandemic flu. Another venue for data aggregation analysis is social media. By examining data aggregated by user posts, researchers are gaining insight into health perceptions and behaviors as well as early detection of potential disease trends. Though criticized early on for the possibility of false reports and lack of specificity and sensitivity, social media's freely available, "real time" and statistically significant data is becoming an essential tool for disease surveillance.⁵²⁻⁵³

Case Study

Mobiles in Malawi was initiated in the summer of 2007, by Josh Nesbitt who was working with a "rural Malawian hospital that serves 250,000 patients spread 100 miles in every direction. To reach remote patients, the hospital trained volunteer community health workers (CHWs) like Dickson Mtanga, a subsistence farmer. Dickson had to walk 35 miles to submit hand-written reports on 25 HIV-positive patients in his community. The hospital needed a simple means of communication."⁴² Seeing the need Josh returned to the hospital the following year with mobile phones and a laptop running *FrontlineSMS*. In late 2008, *Mobiles in Malawi* merged with *MobilizeMRS*, an electronic medical records initiative that trained CHWs in structured data collection. The coming together of these efforts resulted in the creation of *FrontlineSMS:Medic* whose "mission was to help health workers communicate, coordinate patient care, and provide diagnostics using low-cost, appropriate technology...."

"In six months, the pilot in Malawi using *FrontlineSMS* saved hospital staff 1200 hours of follow-up time and over \$3,000 in motorbike fuel. Over 100 patients started tuberculosis treatment after their symptoms were noticed by CHWs and reported by text message. The SMS network brought the Home-Based Care unit to the homes of 130 patients who would not have otherwise received care, and texting saved 21 antiretroviral therapy (ART) monitors 900 hours of travel time, eliminating the need to hand deliver paper reports."⁴⁵

Frontline SMS:Medic has since been deployed in Haiti after the 2010 earthquake where it was used by frontline disaster relief workers to text message urgent needs. "Using crowd-sourced translation, categorization, and geo-tagging, reports were created for first responders within 5 minutes of receiving an SMS. Over 80,000 messages were received in the first five weeks of operation, focusing relief efforts for thousands of Haitians."⁴⁵

"In less than one year, *FrontlineSMS:Medic* expanded from 75 to 1,500 end users linked to clinics serving approximately 3.5 million patients. Growing from the first pilot at a single hospital in Malawi, they established programs in 40% of Malawi's district hospitals and implemented projects in nine other countries, including Honduras, Haiti, Uganda, Mali, Kenya, South Africa, Cameroon, India and Bangladesh."⁴⁵

Frontline SMS has developed other mobile tools including: *PatientView*, a lightweight patient records system, *TextForms*, a text-based information collection module, and a messaging module for *OpenMRS*. *FrontlineSMS:Medic* recently changed its name to *Medic Mobile*.

Key Points

- **Public health informatics** is an **important sub-category of health informatics**
- **Public health reporting** will be part of **meaningful use stages 1-3**
- **Public health surveillance** is very broad and covers **infectious diseases, epidemics, natural disasters and bioterrorism**
- **Geographic information systems** provide a convenient **display of medical information overlaid on geographical interface**
- A **myriad of new national and global public health informatics-related initiatives** have been established

Conclusion

Public health is concerned with the health of populations, instead of individuals. In order to study large populations and track trends in health and other public health activities, paper-based reporting is no longer tenable. A robust public health network will require data standards, electronic health records and health information exchange. As a result of the HITECH Act and Affordable Care Act we are moving closer to the ideal goal of almost real time public health surveillance and reporting.

Acknowledgements

We would like to thank our graduate students for their assistance in preparing this chapter: Sara Beard BS, MPHc and Georgina Palombo MBA.