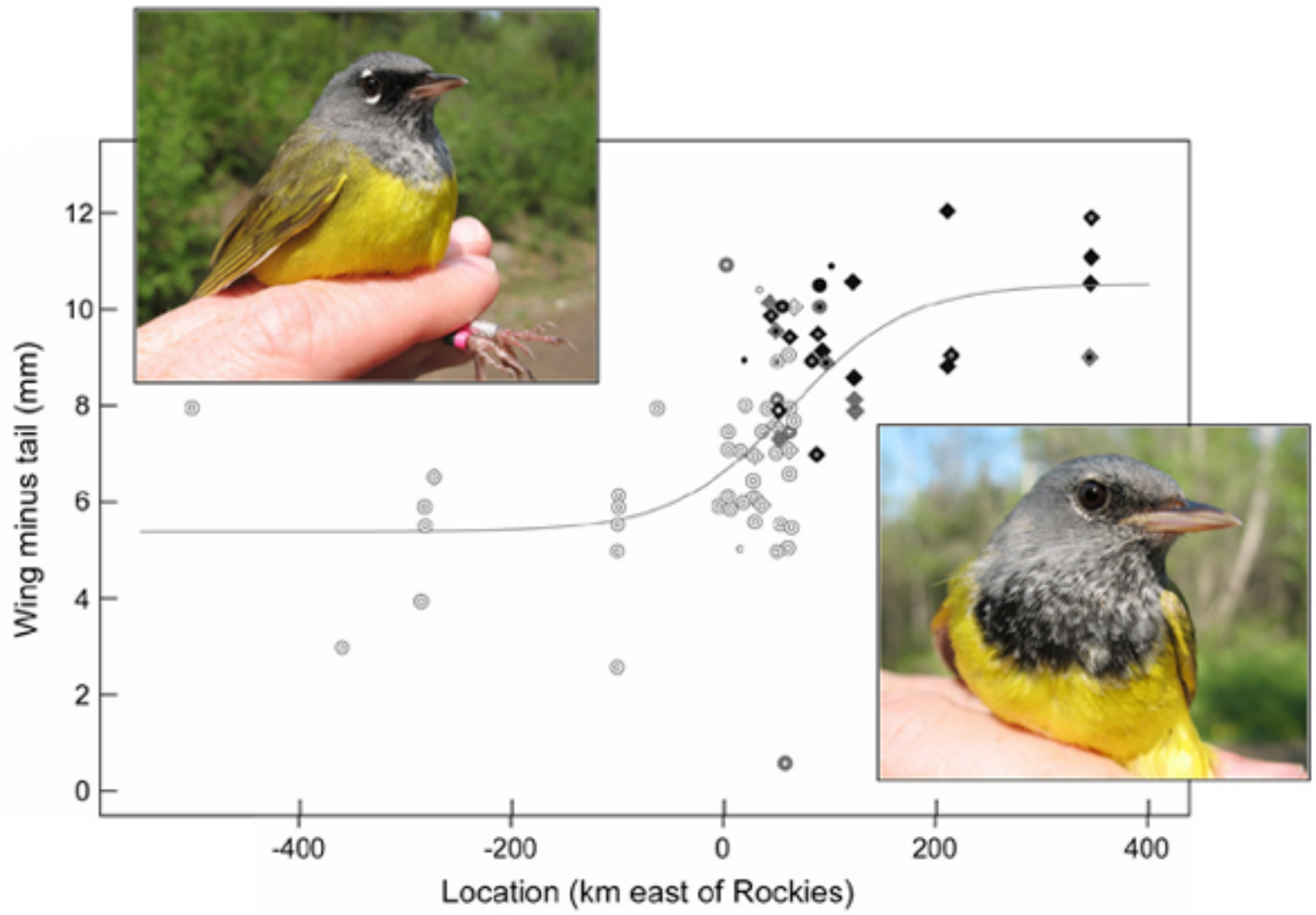
Biostatistics



Katelynn Alcala

First Edition, 2012

ISBN 978-81-323-3249-7

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Introduction

Biostatistics (a contraction of biology and statistics; sometimes referred to as **biometry** or **biometrics**) is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine and agriculture; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.

Biostatistics and the history of biological thought

Biostatistical reasoning and modeling were of critical importance to the foundation theories of modern biology. In the early 1900s, after the rediscovery of Mendel's work, the conceptual gaps in understanding between genetics and evolutionary Darwinism led to vigorous debate between biometricians such as Walter Weldon and Karl Pearson and Mendelians such as Charles Davenport, William Bateson and Wilhelm Johannsen. By the 1930s statisticians and models built on statistical reasoning had helped to resolve these differences and to produce the neo-Darwinian modern evolutionary synthesis.

The leading figures in the establishment of this synthesis all relied on statistics and developed its use in biology.

- Sir Ronald A. Fisher developed several basic statistical methods in support of his work *The Genetical Theory of Natural Selection*
- Sewall G. Wright used statistics in the development of modern population genetics
- J. B. S Haldane's book, *The Causes of Evolution*, reestablished natural selection as the premier mechanism of evolution by explaining it in terms of the mathematical consequences of Mendelian genetics.

These individuals and the work of other biostatisticians, mathematical biologists, and statistically inclined geneticists helped bring together evolutionary biology and genetics into a consistent, coherent whole that could begin to be quantitatively modeled.

In parallel to this overall development, the pioneering work of D'Arcy Thompson in *On Growth and Form* also helped to add quantitative discipline to biological study.

Despite the fundamental importance and frequent necessity of statistical reasoning, there may nonetheless have been a tendency among biologists to distrust or deprecate results which are not qualitatively apparent. One anecdote describes Thomas Hunt Morgan banning the Friden calculator from his department at Caltech, saying "Well, I am like a guy who is prospecting for gold along the banks of the Sacramento River in 1849. With a little intelligence, I can reach down and pick up big nuggets of gold. And as long as I can do that, I'm not going to let any people in my department waste scarce resources in placer mining." Educators are now adjusting their curricula to focus on more quantitative concepts and tools.

Education and training programs

Almost all educational programmes in biostatistics are at postgraduate level. They are most often found in schools of public health, affiliated with schools of medicine, forestry, or agriculture or as a focus of application in departments of statistics.

In the United States, while several universities have dedicated biostatistics departments, many other top-tier universities integrate biostatistics faculty into statistics or other departments, such as epidemiology. Thus departments carrying the name "biostatistics" may exist under quite different structures. For instance, relatively new biostatistics departments have been founded with a focus on bioinformatics and computational biology, whereas older departments, typically affiliated with schools of public health, will have more traditional lines of research involving epidemiological studies and clinical trials as well as bioinformatics. In larger universities where both a statistics and a biostatistics department exist, the degree of integration between the two departments may range from the bare minimum to very close collaboration. In general, the difference between a statistics program and a biostatistics one is twofold: (i) statistics departments will often host theoretical/methodological research which are less common in biostatistics programs and (ii) statistics departments have lines of research that may include biomedical applications but also other areas such as industry (quality control), business and economics and biological areas other than medicine.

Applications of biostatistics

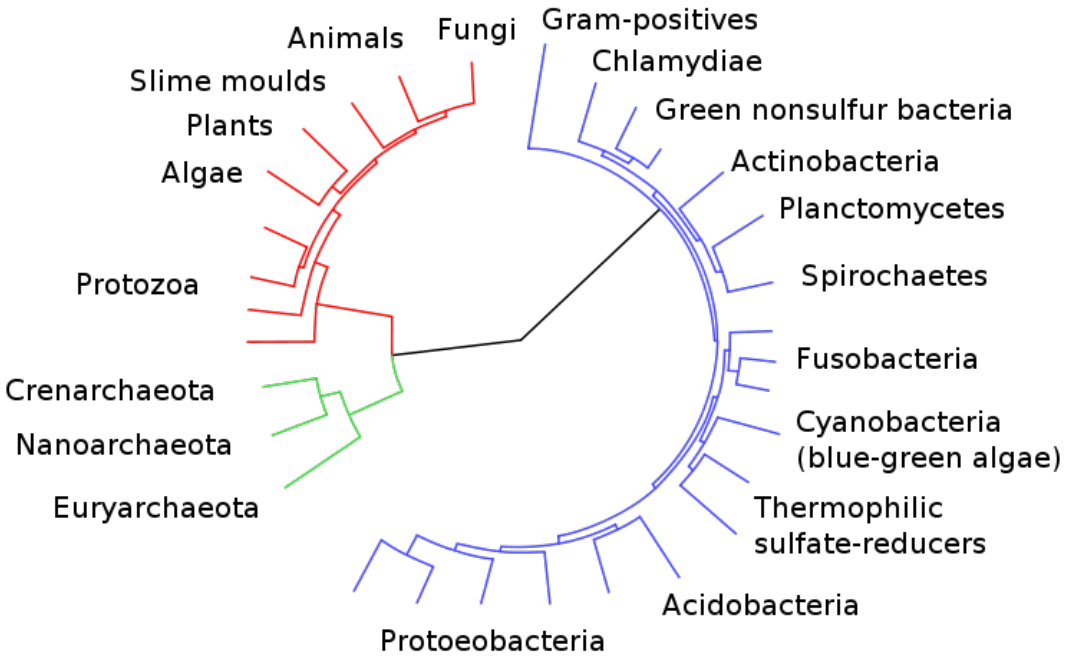
- Public health, including epidemiology, health services research, nutrition, and environmental health
- Design and analysis of clinical trials in medicine
- Population genetics, and statistical genetics in order to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and farm animals (animal breeding). In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics
- Anaysis of genomics data, for example from microarray or proteomics experiments . Often concerning diseases or disease stages.
- Ecology, ecological forecasting
- Biological sequence analysis

- Systems biology for gene network inference or pathways analysis

Statistical methods are beginning to be integrated into medical informatics, public health informatics, bioinformatics and computational biology.

Chapter- 1

Population Genetics



Population genetics is the study of allele frequency distribution and change under the influence of the four main evolutionary processes: natural selection, genetic drift, mutation and gene flow. It also takes into account the factors of population subdivision and population structure. It attempts to explain such phenomena as adaptation and speciation.

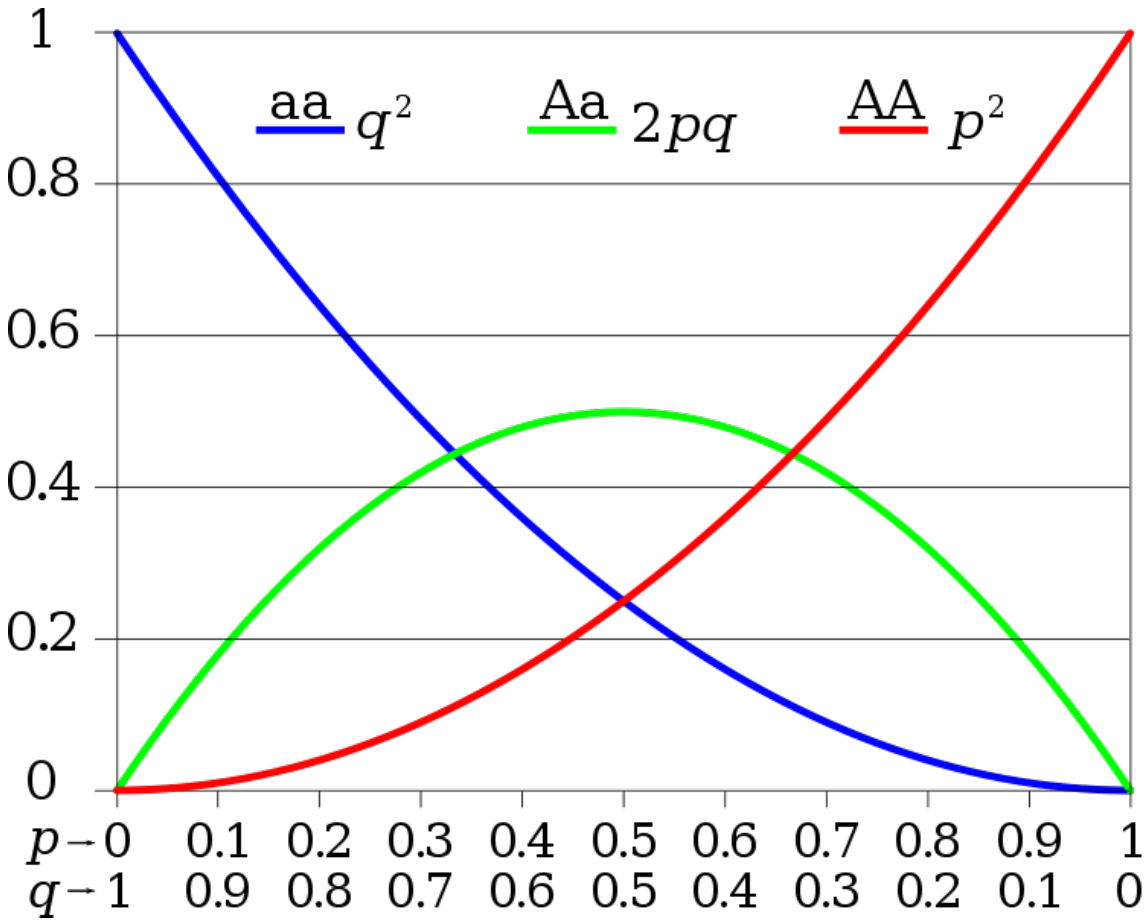
Population genetics was a vital ingredient in the emergence of the modern evolutionary synthesis. Its primary founders were Sewall Wright, J. B. S. Haldane and R. A. Fisher, who also laid the foundations for the related discipline of quantitative genetics.

Fundamentals

Population genetics concerns the genetic constitution of populations and how this constitution changes with time. A population is a set of organisms in which any pair of

members can breed together. This implies that all members belong to the same species and live near each other.

For example, all of the moths of the same species living in an isolated forest are a population. A gene in this population may have several alternate forms, which account for variations between the phenotypes of the organisms. An example might be a gene for coloration in moths that has two alleles: black and white. A gene pool is the complete set of alleles for a gene in a single population; the allele frequency for an allele is the fraction of the genes in the pool that is composed of that allele (for example, what fraction of moth coloration genes are the black allele). Evolution occurs when there are changes in the frequencies of alleles within a population; for example, the allele for black color in a population of moths becoming more common.



Hardy–Weinberg principle for two alleles: the horizontal axis shows the two allele frequencies *p* and *q* and the vertical axis shows the genotype frequencies. Each graph shows one of the three possible genotypes.

Hardy–Weinberg principle

To understand the mechanisms that cause a population to evolve, it is useful to consider what conditions are required for a population not to evolve. The *Hardy-Weinberg*

principle states that the frequencies of alleles (variations in a gene) in a sufficiently large population will remain constant if the only forces acting on that population are the random reshuffling of alleles during the formation of the sperm or egg, and random combination of the alleles in these sex cells during fertilization. Such a population is said to be in *Hardy-Weinberg equilibrium* as it is not evolving. Hardy Weinberg equilibrium is impossible in nature. Genetic equilibrium is an ideal state that provides a baseline to measure genetic change against.

Allele frequencies in a population remain static across generations, provided the following conditions are at hand: random mating, no mutation (the alleles don't change), no migration or emigration (no exchange of alleles between populations), infinitely large population size, and no selective pressure for or against any traits.

In the simplest case of a single locus with two alleles: the dominant allele is denoted **A** and the recessive **a** and their frequencies are denoted by p and q; freq(**A**) = p; freq(**a**) = q; $p + q = 1$. If the population is in equilibrium, then we will have freq(**AA**) = p^2 for the **AA** homozygotes in the population, freq(**aa**) = q^2 for the **aa** homozygotes, and freq(**Aa**) = $2pq$ for the heterozygotes.

Based on these equations, useful but difficult-to-measure facts about a population can be determined. For example, a patient's child is a carrier of a recessive mutation that causes cystic fibrosis in homozygous recessive children. The parent wants to know the probability of her grandchildren inheriting the disease. In order to answer this question, the genetic counselor must know the chance that the child will reproduce with a carrier of the recessive mutation. This fact may not be known, but disease frequency is known. We know that the disease is caused by the homozygous recessive genotype; we can use the Hardy–Weinberg principle to work backward from disease occurrence to the frequency of heterozygous recessive individuals.

Scope and theoretical considerations

The mathematics of population genetics were originally developed as part of the modern evolutionary synthesis. According to Beatty (1986), it defines the core of the modern synthesis.

According to Lewontin (1974), the theoretical task for population genetics is a process in two spaces: a "genotypic space" and a "phenotypic space". The challenge of a *complete* theory of population genetics is to provide a set of laws that predictably map a population of genotypes (G_1) to a phenotype space (P_1), where selection takes place, and another set of laws that map the resulting population (P_2) back to genotype space (G_2) where Mendelian genetics can predict the next generation of genotypes, thus completing the cycle. Even leaving aside for the moment the non-Mendelian aspects of molecular genetics, this is clearly a gargantuan task. Visualizing this transformation schematically:

$$G_1 \xrightarrow{T_1} P_1 \xrightarrow{T_2} P_2 \xrightarrow{T_3} G_2 \xrightarrow{T_4} G_1' \rightarrow \cdots$$

(adapted from Lewontin 1974, p. 12). XD

T_1 represents the genetic and epigenetic laws, the aspects of functional biology, or development, that transform a genotype into phenotype. We will refer to this as the "genotype-phenotype map". T_2 is the transformation due to natural selection, T_3 are epigenetic relations that predict genotypes based on the selected phenotypes and finally T_4 the rules of Mendelian genetics.

In practice, there are two bodies of evolutionary theory that exist in parallel, traditional population genetics operating in the genotype space and the biometric theory used in plant and animal breeding, operating in phenotype space. The missing part is the mapping between the genotype and phenotype space. This leads to a "sleight of hand" (as Lewontin terms it) whereby variables in the equations of one domain, are considered parameters or *constants*, where, in a full-treatment they would be transformed themselves by the evolutionary process and are in reality *functions* of the state variables in the other domain. The "sleight of hand" is assuming that we know this mapping. Proceeding as if we do understand it is enough to analyze many cases of interest. For example, if the phenotype is almost one-to-one with genotype (sickle-cell disease) or the time-scale is sufficiently short, the "constants" can be treated as such; however, there are many situations where it is inaccurate.

The four processes

Natural selection

Natural selection is the process by which heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations.

The natural genetic variation within a population of organisms means that some individuals will survive more successfully than others in their current environment. Factors which affect reproductive success are also important, an issue which Charles Darwin developed in his ideas on sexual selection.

Natural selection acts on the phenotype, or the observable characteristics of an organism, but the genetic (heritable) basis of any phenotype which gives a reproductive advantage will become more common in a population. Over time, this process can result in adaptations that specialize organisms for particular ecological niches and may eventually result in the emergence of new species.

Natural selection is one of the cornerstones of modern biology. The term was introduced by Darwin in his groundbreaking 1859 book *On the Origin of Species*, in which natural selection was described by analogy to artificial selection, a process by which animals and plants with traits considered desirable by human breeders are systematically favored for reproduction. The concept of natural selection was originally developed in the absence of a valid theory of heredity; at the time of Darwin's writing, nothing was known of modern

genetics. The union of traditional Darwinian evolution with subsequent discoveries in classical and molecular genetics is termed the *modern evolutionary synthesis*. Natural selection remains the primary explanation for adaptive evolution.

Genetic drift

Genetic drift is the change in the relative frequency in which a gene variant (allele) occurs in a population due to random sampling and chance. That is, the alleles in the offspring in the population are a random sample of those in the parents. And chance has a role in determining whether a given individual survives and reproduces. A population's allele frequency is the fraction or percentage of its gene copies compared to the total number of gene alleles that share a particular form.

Genetic drift is an important evolutionary process which leads to changes in allele frequencies over time. It may cause gene variants to disappear completely, and thereby reduce genetic variability. In contrast to natural selection, which makes gene variants more common or less common depending on their reproductive success, the changes due to genetic drift are not driven by environmental or adaptive pressures, and may be beneficial, neutral, or detrimental to reproductive success.

The effect of genetic drift is larger in small populations, and smaller in large populations. Vigorous debates wage among scientists over the relative importance of genetic drift compared with natural selection. Ronald Fisher held the view that genetic drift plays at the most a minor role in evolution, and this remained the dominant view for several decades. In 1968 Motoo Kimura rekindled the debate with his neutral theory of molecular evolution which claims that most of the changes in the genetic material are caused by genetic drift.

Mutation

Mutations are changes in the DNA sequence of a cell's genome and are caused by radiation, viruses, transposons and mutagenic chemicals, as well as errors that occur during meiosis or DNA replication. Errors are introduced particularly often in the process of DNA replication, in the polymerization of the second strand. These errors can also be induced by the organism itself, by cellular processes such as hypermutation.

Mutations can have an impact on the phenotype of an organism, especially if they occur within the protein coding sequence of a gene. Error rates are usually very low (1 error in every 10 million–100 million bases) due to the "proofreading" ability of DNA polymerases. Without proofreading, error rates are a thousandfold higher. Chemical damage to DNA occurs naturally as well, and cells use DNA repair mechanisms to repair mismatches and breaks in DNA. Nevertheless, the repair sometimes fails to return the DNA to its original sequence.

In organisms that use chromosomal crossover to exchange DNA and recombine genes, errors in alignment during meiosis can also cause mutations. Errors in crossover are

especially likely when similar sequences cause partner chromosomes to adopt a mistaken alignment; this makes some regions in genomes more prone to mutating in this way. These errors create large structural changes in DNA sequence—duplications, inversions or deletions of entire regions, or the accidental exchanging of whole parts between different chromosomes (called translocation).

Mutation can result in several different types of change in DNA sequences; these can either have no effect, alter the product of a gene, or prevent the gene from functioning. Studies in the fly *Drosophila melanogaster* suggest that if a mutation changes a protein produced by a gene, this will probably be harmful, with about 70 percent of these mutations having damaging effects, and the remainder being either neutral or weakly beneficial. Due to the damaging effects that mutations can have on cells, organisms have evolved mechanisms such as DNA repair to remove mutations. Therefore, the optimal mutation rate for a species is a trade-off between costs of a high mutation rate, such as deleterious mutations, and the metabolic costs of maintaining systems to reduce the mutation rate, such as DNA repair enzymes. Viruses that use RNA as their genetic material have rapid mutation rates, which can be an advantage since these viruses will evolve constantly and rapidly, and thus evade the defensive responses of e.g. the human immune system.

Mutations can involve large sections of DNA becoming duplicated, usually through genetic recombination. These duplications are a major source of raw material for evolving new genes, with tens to hundreds of genes duplicated in animal genomes every million years. Most genes belong to larger families of genes of shared ancestry. Novel genes are produced by several methods, commonly through the duplication and mutation of an ancestral gene, or by recombining parts of different genes to form new combinations with new functions.

Here, domains act as modules, each with a particular and independent function, that can be mixed together to produce genes encoding new proteins with novel properties. For example, the human eye uses four genes to make structures that sense light: three for color vision and one for night vision; all four arose from a single ancestral gene. Another advantage of duplicating a gene (or even an entire genome) is that this increases redundancy; this allows one gene in the pair to acquire a new function while the other copy performs the original function. Other types of mutation occasionally create new genes from previously noncoding DNA.

Gene flow

Gene flow is the exchange of genes between populations, which are usually of the same species. Examples of gene flow within a species include the migration and then breeding of organisms, or the exchange of pollen. Gene transfer between species includes the formation of hybrid organisms and horizontal gene transfer.

Migration into or out of a population can change allele frequencies, as well as introducing genetic variation into a population. Immigration may add new genetic material to the

established gene pool of a population. Conversely, emigration may remove genetic material. As barriers to reproduction between two diverging populations are required for the populations to become new species, gene flow may slow this process by spreading genetic differences between the populations. Gene flow is hindered by mountain ranges, oceans and deserts or even man-made structures such as the Great Wall of China, which has hindered the flow of plant genes.

Depending on how far two species have diverged since their most recent common ancestor, it may still be possible for them to produce offspring, as with horses and donkeys mating to produce mules. Such hybrids are generally infertile, due to the two different sets of chromosomes being unable to pair up during meiosis. In this case, closely related species may regularly interbreed, but hybrids will be selected against and the species will remain distinct. However, viable hybrids are occasionally formed and these new species can either have properties intermediate between their parent species, or possess a totally new phenotype. The importance of hybridization in creating new species of animals is unclear, although cases have been seen in many types of animals, with the gray tree frog being a particularly well-studied example.

Hybridization is, however, an important means of speciation in plants, since polyploidy (having more than two copies of each chromosome) is tolerated in plants more readily than in animals. Polyploidy is important in hybrids as it allows reproduction, with the two different sets of chromosomes each being able to pair with an identical partner during meiosis. Polyploids also have more genetic diversity, which allows them to avoid inbreeding depression in small populations.

Horizontal gene transfer is the transfer of genetic material from one organism to another organism that is not its offspring; this is most common among bacteria. In medicine, this contributes to the spread of antibiotic resistance, as when one bacteria acquires resistance genes it can rapidly transfer them to other species. Horizontal transfer of genes from bacteria to eukaryotes such as the yeast *Saccharomyces cerevisiae* and the adzuki bean beetle *Callosobruchus chinensis* may also have occurred. An example of larger-scale transfers are the eukaryotic bdelloid rotifers, which appear to have received a range of genes from bacteria, fungi, and plants. Viruses can also carry DNA between organisms, allowing transfer of genes even across biological domains. Large-scale gene transfer has also occurred between the ancestors of eukaryotic cells and prokaryotes, during the acquisition of chloroplasts and mitochondria.

Gene flow is the transfer of alleles from one population to another.

Migration into or out of a population may be responsible for a marked change in allele frequencies. Immigration may also result in the addition of new genetic variants to the established gene pool of a particular species or population.

There are a number of factors that affect the rate of gene flow between different populations. One of the most significant factors is mobility, as greater mobility of an

individual tends to give it greater migratory potential. Animals tend to be more mobile than plants, although pollen and seeds may be carried great distances by animals or wind.

Maintained gene flow between two populations can also lead to a combination of the two gene pools, reducing the genetic variation between the two groups. It is for this reason that gene flow strongly acts against speciation, by recombining the gene pools of the groups, and thus, repairing the developing differences in genetic variation that would have led to full speciation and creation of daughter species.

For example, if a species of grass grows on both sides of a highway, pollen is likely to be transported from one side to the other and vice versa. If this pollen is able to fertilise the plant where it ends up and produce viable offspring, then the alleles in the pollen have effectively been able to move from the population on one side of the highway to the other.

Genetic structure

Because of physical barriers to migration, along with limited vagility, and natal philopatry, natural populations are rarely panmictic (Buston *et al.*, 2007). There is usually a geographic range within which individuals are more closely related to one another than those randomly selected from the general population. This is described as the extent to which a population is genetically structured (Repaci *et al.*, 2007).

Microbial population genetics

Microbial population genetics is a rapidly advancing field of investigation with relevance to many other theoretical and applied areas of scientific investigations. The population genetics of microorganisms lays the foundations for tracking the origin and evolution of antibiotic resistance and deadly infectious pathogens. Population genetics of microorganisms is also an essential factor for devising strategies for the conservation and better utilization of beneficial microbes (Xu, 2010).

History

Biston betularia f. typica is the white-bodied form of the peppered moth.

Biston betularia f. carbonaria is the black-bodied form of the peppered moth.

Population genetics

Population genetics was developed as a reconciliation of the Mendelian and biometrician models. A key step was the work of the British biologist and statistician R.A. Fisher. In a series of papers starting in 1918 and culminating in his 1930 book *The Genetical Theory of Natural Selection*, Fisher showed that the continuous variation measured by the biometricians could be produced by the combined action of many discrete genes, and that natural selection could change gene frequencies in a population, resulting in evolution (though lacking the knowledge of what an actual gene was at this time, it should be said in this sense he understood phenotypic trait frequency, rather than specifically identifiable gene frequency). In a series of papers beginning in 1924, another British geneticist, J.B.S. Haldane, applied statistical analysis to real-world examples of natural selection, such as the evolution of industrial melanism in peppered moths, and showed that natural selection worked at an even faster rate than Fisher assumed.

The American biologist Sewall Wright, who had a background in animal breeding experiments, focused on combinations of interacting genes, and the effects of inbreeding on small, relatively isolated populations that exhibited genetic drift. In 1932, Wright introduced the concept of an adaptive landscape and argued that genetic drift and inbreeding could drive a small, isolated sub-population away from an adaptive peak, allowing natural selection to drive it towards different adaptive peaks. Fisher and Wright had some fundamental disagreements and a controversy about the relative roles of selection and drift continued for much of the century between the Americans and the British. The Frenchman Gustave Malécot was also important early in the development of the discipline.

The work of Fisher, Haldane and Wright founded the discipline of *population genetics*. This integrated natural selection with Mendelian genetics, which was the critical first step in developing a unified theory of how evolution worked.

John Maynard Smith was Haldane's pupil, whilst W.D. Hamilton was heavily influenced by the writings of Fisher. The American George R. Price worked with both Hamilton and Maynard Smith. American Richard Lewontin and Japanese Motoo Kimura were heavily influenced by Wright.

Modern evolutionary synthesis

In the first few decades of the 20th century, most field naturalists continued to believe that Lamarckian and orthogenic mechanisms of evolution provided the best explanation for the complexity they observed in the living world. However, as the field of genetics continued to develop, those views became less tenable. Theodosius Dobzhansky, a postdoctoral worker in T. H. Morgan's lab, had been influenced by the work on genetic diversity by Russian geneticists such as Sergei Chetverikov. He helped to bridge the divide between the foundations of microevolution developed by the population geneticists and the patterns of macroevolution observed by field biologists, with his 1937 book *Genetics and the Origin of Species*.

Dobzhansky examined the genetic diversity of wild populations and showed that, contrary to the assumptions of the population geneticists, these populations had large amounts of genetic diversity, with marked differences between sub-populations. The book also took the highly mathematical work of the population geneticists and put it into a more accessible form. In Great Britain E.B. Ford, the pioneer of ecological genetics, continued throughout the 1930s and 1940s to demonstrate the power of selection due to ecological factors including the ability to maintain genetic diversity through genetic polymorphisms such as human blood types. Ford's work would contribute to a shift in emphasis during the course of the modern synthesis towards natural selection over genetic drift.

Chapter- 2

Ecological Forecasting and Systems Biology

Ecological forecasting

Ecological forecasting uses knowledge of physics, ecology and physiology to predict how ecosystems will change in the future in response to environmental factors such as climate change. The ultimate goal of the approach is to provide people such as resource managers and designers of marine reserves with information that they can then use to respond, in advance, to future changes, a form of adaptation to global warming.

One of the most important environmental factors impacting organisms today is global warming. Most physiological processes are affected by temperature, and so even small changes in weather and climate can lead to large changes in the growth, reproduction and survival of animals and plants. The scientific consensus is that the increase in atmospheric greenhouse gases due to human activity caused most of the warming observed since the start of the industrial era. These changes are in turn impacting both humans and natural ecosystems.

One major challenge is to predict where, when and with what magnitude impacts are likely to occur so that we can mitigate or at least prepare for them. Ecological forecasting applies existing knowledge of how animals and plants interact with their physical environment to ask how changes in environmental factors might result in changes to the ecosystems as a whole.

Approaches

- **Palaeobiology modeling**: uses fossil and phylogenetic evidence of biodiversity in the past to project the trajectory of biodiversity in the future. Simple plots can be constructed and then adjusted based on the varying quality of the fossil record.

- **Climate envelope modeling**: relies on statistical correlations between existing species distributions and environmental variables to define a species' tolerance. *Envelopes* of tolerance are then drawn around existing ranges. By predicting future levels of factors such as temperature, rainfall, and salinity, new range boundaries are then predicted. These methods are good for examining large numbers of species, but are likely not a good means of predicting effects at fine scales.

- **Niche level modeling**: is a newer method which links physiological information about a species to models of animal and plant body temperature. In contrast to "climate envelope" approaches, environmental variables are predicted at the level of the niche and are therefore much more exact. However, the approach is also usually more time consuming.

Forecasting examples

Biodiversity

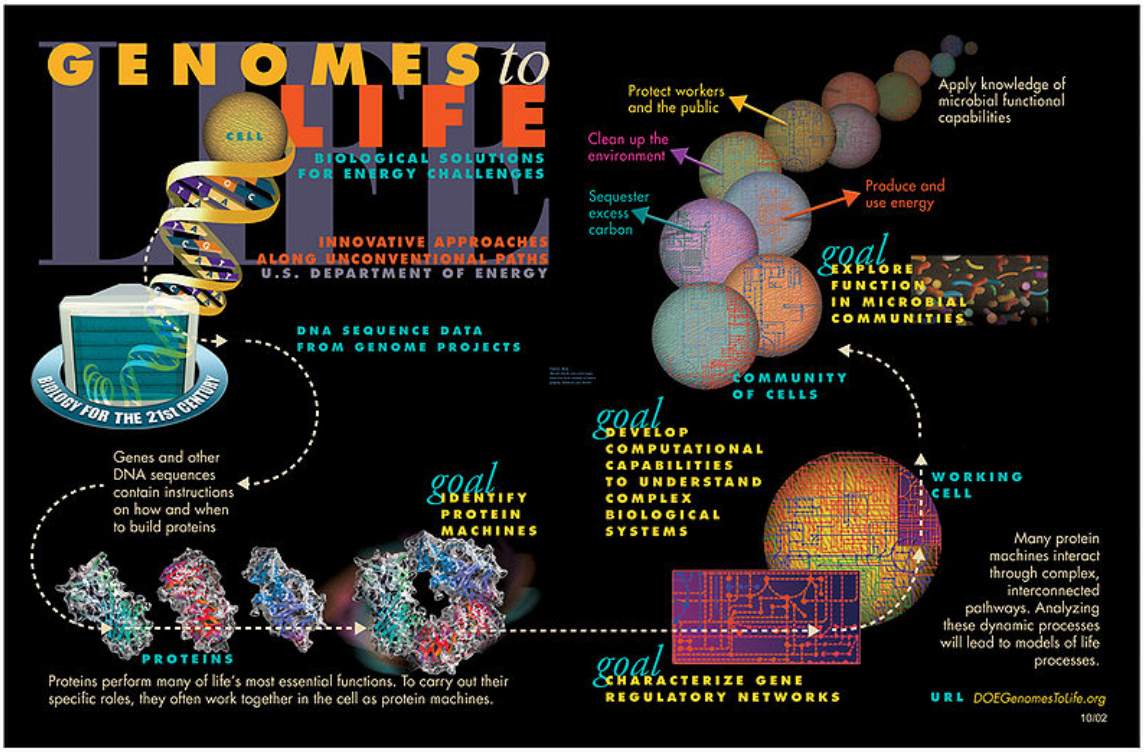
Using fossil evidence, studies have shown that vertebrate biodiversity has grown exponentially through Earth's history and that biodiversity is entwined with the diversity of Earth's habitats.

"Animals have not yet invaded 2/3 of Earth's habitats, and it could be that without human influence biodiversity will continue to increase in an exponential fashion."
—Sahney *et al*.

Temperature

Forecasts of temperature, shown in the diagram at the right as colored dots, along the North Island of New Zealand in the austral summer of 2007. As per the temperature scale shown at the bottom, intertidal temperatures were forecast to exceed 30°C at some locations on February 19; surveys later showed that these sites corresponded to large die-offs in burrowing sea urchins.

Systems biology



Example of systems biology research

Systems biology is a term used to describe a number of trends in bioscience research, and a movement which draws on those trends. Proponents describe systems biology as a biology-based inter-disciplinary study field that focuses on complex interactions in biological systems, claiming that it uses a new perspective (holism instead of reduction). Particularly from year 2000 onwards, the term is used widely in the biosciences, and in a variety of contexts. An often stated ambition of systems biology is the modeling and discovery of emergent properties, properties of a system whose theoretical description is only possible using techniques which fall under the remit of systems biology.

Overview

Systems biology can be considered from a number of different aspects:

- As a **field of study**, particularly, the study of the interactions between the components of *biological systems*, and how these interactions give rise to the function and behavior of that system (for example, the enzymes and metabolites in a metabolic pathway).

- As a **paradigm**, usually defined in antithesis to the so-called reductionist paradigm (biological organisation), although fully consistent with the scientific

method. The distinction between the two paradigms is referred to in these quotations:

"The reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge...the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models" Science
"Systems biology...is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different....It means changing our philosophy, in the full sense of the term" Denis Noble

- As a series of **operational protocols used for performing research**, namely a cycle composed of theory, analytic or computational modelling to propose specific testable hypotheses about a biological system, experimental validation, and then using the newly acquired quantitative description of cells or cell processes to refine the computational model or theory. Since the objective is a model of the interactions in a system, the experimental techniques that most suit systems biology are those that are system-wide and attempt to be as complete as possible. Therefore, transcriptomics, metabolomics, proteomics and high-throughput techniques are used to collect quantitative data for the construction and validation of models.

- As the application of dynamical systems theory to molecular biology.

- As a **socioscientific phenomenon** defined by the strategy of pursuing integration of complex data about the interactions in biological systems from diverse experimental sources using interdisciplinary tools and personnel.

This variety of viewpoints is illustrative of the fact that systems biology refers to a cluster of peripherally overlapping concepts rather than a single well-delineated field. However the term has widespread currency and popularity as of 2007, with chairs and institutes of systems biology proliferating worldwide.

History

Systems biology finds its roots in:

- the quantitative modeling of enzyme kinetics, a discipline that flourished between 1900 and 1970,
- the mathematical modeling of population growth,
- the simulations developed to study neurophysiology, and
- control theory and cybernetics.

One of the theorists who can be seen as one of the precursors of systems biology is Ludwig von Bertalanffy with his general systems theory. One of the first numerical simulations in biology was published in 1952 by the British neurophysiologists and Nobel prize winners Alan Lloyd Hodgkin and Andrew Fielding Huxley, who constructed a mathematical model that explained the action potential propagating along the axon of a neuronal cell. Their model described a cellular function emerging from the interaction between two different molecular components, a potassium and a sodium channels, and can therefore be seen as the beginning of computational systems biology. In 1960, Denis Noble developed the first computer model of the heart pacemaker.

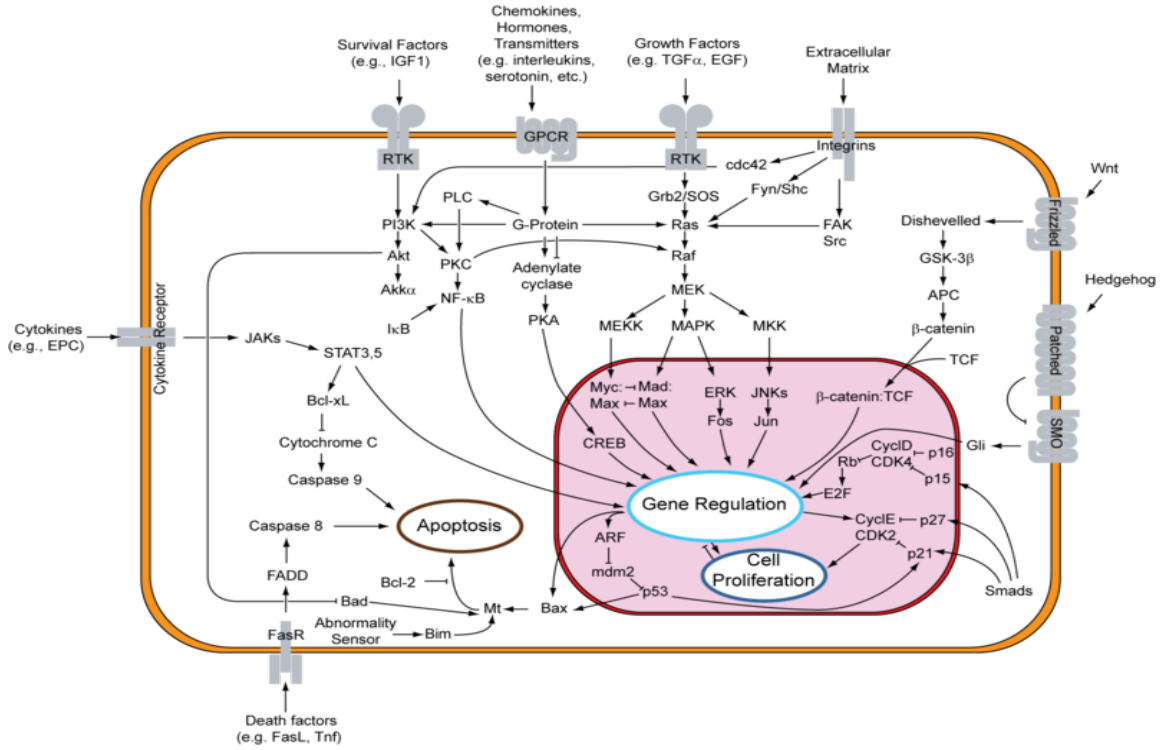
The formal study of systems biology, as a distinct discipline, was launched by systems theorist Mihajlo Mesarovic in 1966 with an international symposium at the Case Institute of Technology in Cleveland, Ohio entitled "Systems Theory and Biology."

The 1960s and 1970s saw the development of several approaches to study complex molecular systems, such as the Metabolic Control Analysis and the biochemical systems theory. The successes of molecular biology throughout the 1980s, coupled with a skepticism toward theoretical biology, that then promised more than it achieved, caused the quantitative modelling of biological processes to become a somewhat minor field.

However the birth of functional genomics in the 1990s meant that large quantities of high quality data became available, while the computing power exploded, making more realistic models possible. In 1997, the group of Masaru Tomita published the first quantitative model of the metabolism of a whole (hypothetical) cell.

Around the year 2000, after Institutes of Systems Biology were established in Seattle and Tokyo, systems biology emerged as a movement in its own right, spurred on by the completion of various genome projects, the large increase in data from the omics (e.g. genomics and proteomics) and the accompanying advances in high-throughput experiments and bioinformatics. Since then, various research institutes dedicated to systems biology have been developed. As of summer 2006, due to a shortage of people in systems biology several doctoral training centres in systems biology have been established in many parts of the world.

Disciplines associated with systems biology



Overview of signal transduction pathways

According to the interpretation of Systems Biology as the ability to obtain, integrate and analyze complex data from multiple experimental sources using interdisciplinary tools, some typical technology platforms are:

- Phenomics: Organismal variation in phenotype as it changes during its life span.
- Genomics: Organismal deoxyribonucleic acid (DNA) sequence, including intra-organisamal cell specific variation. (i.e. Telomere length variation etc.).
- Epigenomics / Epigenetics: Organismal and corresponding cell specific transcriptomic regulating factors not empirically coded in the genomic sequence. (i.e. DNA methylation, Histone Acetelation etc.).
- Transcriptomics: Organismal, tissue or whole cell gene expression measurements by DNA microarrays or serial analysis of gene expression
- Interferomics: Organismal, tissue, or cell level transcript correcting factors (i.e. RNA interference)
- Translatomics / Proteomics: Organismal, tissue, or cell level measurements of proteins and peptides via two-dimensional gel electrophoresis, mass spectrometry or multi-dimensional protein identification techniques (advanced HPLC systems coupled with mass spectrometry). Sub disciplines include phosphoproteomics, glycoproteomics and other methods to detect chemically modified proteins.
- Metabolomics: Organismal, tissue, or cell level measurements of all small-molecules known as metabolites.
- Glycomics: Organismal, tissue, or cell level measurements of carbohydrates.

- Lipidomics: Organismal, tissue, or cell level measurements of lipids.

In addition to the identification and quantification of the above given molecules further techniques analyze the dynamics and interactions within a cell. This includes:

- Interactomics: Organismal, tissue, or cell level study of interactions between molecules. Currently the authoritative molecular discipline in this field of study is protein-protein interactions (PPI), although the working definition does not preclude inclusion of other molecular disciplines such as those defined here.
- Fluxomics: Organismal, tissue, or cell level measurements of molecular dynamic changes over time.
- Biomics: systems analysis of the biome.

The investigations are frequently combined with large scale perturbation methods, including gene-based (RNAi, mis-expression of wild type and mutant genes) and chemical approaches using small molecule libraries. Robots and automated sensors enable such large-scale experimentation and data acquisition. These technologies are still emerging and many face problems that the larger the quantity of data produced, the lower the quality. A wide variety of quantitative scientists (computational biologists, statisticians, mathematicians, computer scientists, engineers, and physicists) are working to improve the quality of these approaches and to create, refine, and retest the models to accurately reflect observations.

The systems biology approach often involves the development of mechanistic models, such as the reconstruction of dynamic systems from the quantitative properties of their elementary building blocks. For instance, a cellular network can be modelled mathematically using methods coming from chemical kinetics and control theory. Due to the large number of parameters, variables and constraints in cellular networks, numerical and computational techniques are often used.

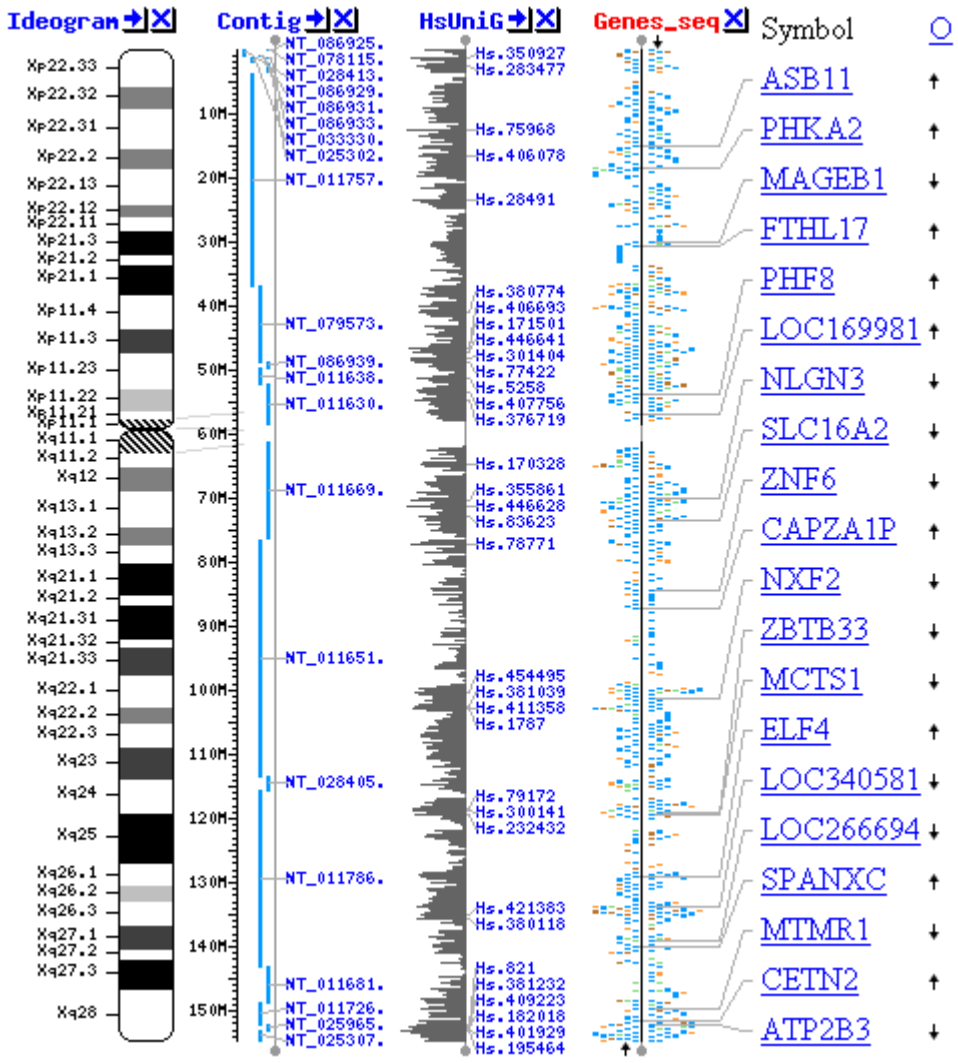
Other aspects of computer science and informatics are also used in systems biology. These include:

- New forms of computational model, such as the use of process calculi to model biological processes (notable approaches include stochastic π-calculus, BioAmbients, Beta Binders, BioPEPA and Brane calculus) and constraint-based modeling.
- Integration of information from the literature, using techniques of information extraction and text mining.
- Development of online databases and repositories for sharing data and models, approaches to database integration and software interoperability via loose coupling of software, websites and databases, or commercial suits.
- Development of syntactically and semantically sound ways of representing biological models.

Chapter- 3

Bioinformatics and Public Health Informatics

Bioinformatics



Map of the human X chromosome. Assembly of the human genome is one of the greatest achievements of bioinformatics.

Bioinformatics is the application of statistics and computer science to the field of molecular biology.

The term *bioinformatics* was coined by Paulien Hogeweg and Ben Hesper in 1978 for the study of informatic processes in biotic systems. Its primary use since at least the late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. It is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization) to achieve this goal. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, genome-wide association studies and the modeling of evolution.

Introduction

Bioinformatics was applied in the creation and maintenance of a database to store biological information at the beginning of the "genomic revolution", such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data.

In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- the development and implementation of tools that enable efficient access to, and use and management of, various types of information.
- the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.

There are two fundamental ways of modelling a Biological system (e.g. living cell) both coming under Bioinformatic approaches.

- Static
    - Sequences - Proteins, Nucleic acids and Peptides
    - Structures - Proteins, Nucleic acids, Ligands (including metabolites and drugs) and Peptides
    - Interaction data among the above entities including microarray data and Networks of proteins, metabolites
- Dynamic
    - Systems Biology comes under this category including reaction fluxes and variable concentrations of metabolites
    - Multi-Agent Based modelling approaches capturing cellular events such as signalling, transcription and reaction dynamics

A broad sub-category under bioinformatics is structural bioinformatics.

Major research areas

Sequence analysis

Since the Phage Φ-X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs such as BLAST are used daily to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides. These programs can compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. A variant of this sequence alignment is used in the sequencing process itself. The so-called shotgun sequencing technique (which was used, for example, by The Institute for Genomic Research to sequence the first bacterial genome, *Haemophilus influenzae*) does not produce entire chromosomes, but instead generates the sequences of many thousands of small DNA fragments (ranging from 35 to 900 nucleotides long, depending on the sequencing technology). The ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome.

Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly will usually contain numerous gaps that have to be filled in later. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

Another aspect of bioinformatics in sequence analysis is annotation, which involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome. Not all of the nucleotides within a genome are part of genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects — for example, in the use of DNA sequences for protein identification.

Genome annotation

In the context of genomics, **annotation** is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White, who was part of the team at The Institute for Genomic Research that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,
- build complex computational models of populations to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

Future work endeavours to reconstruct the now more complex tree of life.

The area of research within computer science that uses genetic algorithms is sometimes confused with computational evolutionary biology, but the two areas are not necessarily related.

Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

Analysis of regulation

Regulation is the complex orchestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins. Bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene. These motifs influence the extent to which that region is transcribed into mRNA. Expression data can be used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). One can then apply clustering algorithms to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented regulatory elements.

Analysis of protein expression

Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

Analysis of mutations in cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms. New physical detection technologies are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single-nucleotide polymorphism arrays to detect known *point mutations*. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or noise, and thus Hidden Markov model and change-point analysis methods are being developed to infer real copy number changes.

Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumors.

Comparative genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

Many of these studies are based on the homology detection and protein families computation.

Modeling biological systems

Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the

complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

High-throughput image analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical imagery. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed. A fully developed analysis system may completely replace the observer. Although these systems are not unique to biomedical imagery, biomedical imaging is becoming more important for both diagnostics and research. Some examples are:

- high-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology, Bioimage informatics)
- morphometrics
- clinical image analysis and visualization
- determining the real-time air-flow patterns in breathing lungs of living animals
- quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
- making behavioral observations from extended video recordings of laboratory animals
- infrared measurements for metabolic activity determination
- inferring clone overlaps in DNA mapping, e.g. the Sulston score

Structural Bioinformatic Approaches

Prediction of protein structure

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the bovine spongiform encephalopathy - aka Mad Cow Disease - prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary*, *tertiary* and *quaternary* structure. A viable general solution to such predictions remains an open problem. As of now, most efforts have been directed towards heuristics that work most of the time.

One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B,* whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are

important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably.

One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.

Molecular Interaction

Efficient software is available today for studying interactions among proteins, ligands and peptides. Types of interactions most often encountered in the field include - Protein-ligand (including drug), protein-protein and protein-peptide.

Molecular dynamic simulation of movement of atoms about rotatable bonds is the fundamental principle behind computational algorithms, termed **docking algorithms** for studying molecular interactions.

Docking algorithms

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much work to be done in this field.

Software and tools

Software tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services available from various bioinformatics companies or public institutions.

Web services in bioinformatics

SOAP and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of

the world. The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads.

Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment) and BSA (Biological Sequence Analysis). The availability of these service-oriented bioinformatics resources demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

Public health informatics

Public Health Informatics has been defined as the systematic application of information and computer science and technology to public health practice, research, and learning. It is one of the subdomains of health informatics.

United States

In the United States, public health informatics is practiced by individuals in public health agencies at the federal and state levels and in the larger local health jurisdictions. Additionally, research and training in public health informatics takes place at a variety of academic institutions.

At the federal Centers for Disease Control and Prevention in Atlanta, Georgia, the Public Health Informatics and Technology Program Office (PHITPO) focuses on advancing the state of information science and applies digital information technologies to aid in the detection and management of diseases and syndromes in individuals and populations. The three sub-units within PHITPO include Informatics Practice, Policy and Coordination; Informatics Solutions and Operations; and Informatics Research and Development.

The bulk of the work of public health informatics in the United States, as with public health generally, takes place at the state and local level, in the state departments of health and the county or parish departments of health. At a state health department the activities may include: collection and storage of *vital statistics* (birth and death records); collection of reports of communicable disease cases from doctors, hospitals, and laboratories, used for infectious disease surveillance; display of infectious disease statistics and trends; collection of child immunization and lead screening information; daily collection and analysis of emergency room data to detect early evidence of biological threats; collection of hospital capacity information to allow for planning of responses in case of emergencies. Each of these activities presents its own information processing challenge.

Collection of public health data

Before the advent of the internet, public health data in the United States, like other healthcare and business data, were collected on paper forms and stored centrally at the relevant public health agency. If the data were to be computerized they required a distinct data entry process, were stored in the various file formats of the day and analyzed by mainframe computers using standard batch processing.

(TODO: describe CDC-provided DOS/desktop-based systems like TIMSS (TB), STDMIS (Sexually transmitted diseases); Epi-Info for epidemiology investigations; and others)

Since the beginning of the World Wide Web, public health agencies with sufficient information technology resources have been transitioning to web-based collection of public health data, and, more recently, to automated messaging of the same information. In the years roughly 2000 to 2005 the Centers for Disease Control and Prevention, under its National Electronic Disease Surveillance System (NEDSS), built and provided free to states a comprehensive web and message-based reporting system called the NEDSS Base System (NBS). Many states and even larger counties have built their own versions of electronic disease surveillance systems, such as Pennsylvania's PA-NEDSS.

To promote interoperability, the CDC has encouraged the adoption in public health data exchange of several standard vocabularies and messaging formats from the health care world. The most prominent of these are: the Health Level 7 (HL7) standards for health care messaging; the LOINC system for encoding laboratory test and result information; and the Systematized Nomenclature of Medicine (SNOMED) vocabulary of health care concepts.

Since about 2005, the CDC has promoted the idea of the Public Health Information Network to facilitate the transmission of data from various partners in the health care industry and elsewhere (hospitals, clinical and environmental laboratories, doctors' practices, pharmacies) to local health agencies, then to state health agencies, and then to the CDC. At each stage the entity must be capable of receiving the data, storing it, aggregating it appropriately, and transmitting it to the next level. A typical example would be infectious disease data, which hospitals, labs, and doctors are legally required to report to local health agencies; local health agencies must report to their state public health department; and which the states must report in aggregate form to the CDC. Among other uses, the CDC publishes the Morbidity and Mortality Weekly Report (MMWR) based on these data acquired systematically from across the United States.

Major issues in the collection of public health data are: awareness of the need to report data; lack of resources of either the reporter or collector; lack of interoperability of data interchange formats, which can be at the purely syntactic or at the semantic level; variation in reporting requirements across the states, territories, and localities.

Storage of public health data

Storage of public health data shares the same data management issues as other industries. And like other industries, the details of how these issues play out are affected by the nature of the data being managed.

Due to the complexity and variability of public health data, like health care data generally, the issue of data modeling presents a particular challenge. While a generation ago flat data sets for statistical analysis were the norm, today's requirements of interoperability and integrated sets of data across the public health enterprise require more sophistication. The relational database is increasingly the norm in public health informatics. Designers and implementers of the many sets of data required for various public health purposes must find a workable balance between very complex and abstract data models such as HL7's Reference Information Model (RIM) or CDC's Public Health Logical Data Model, and simplistic, ad hoc models that untrained public health practitioners come up with and feel capable of working with.

Due to the variability of the incoming data to public health jurisdictions, data quality assurance is also a major issue.

Analysis of public health data

The need to extract usable public health information from the mass of data available requires the public health informaticist to become familiar with a range of analysis tools, ranging from business intelligence tools to produce routine or ad hoc reports, to sophisticated statistical analysis tools such as DAP/SAS and PSPP/SPSS, to Geographical Information Systems (GIS) to expose the geographical dimension of public health trends.

Applications in health surveillance and epidemiology

- SAPPHIRE (Health care) or *Situational Awareness and Preparedness for Public Health Incidences and Reasoning Engines* is a semantics-based health information system capable of tracking and evaluating situations and occurrences that may affect public health.

Chapter- 4

Group Size Measures

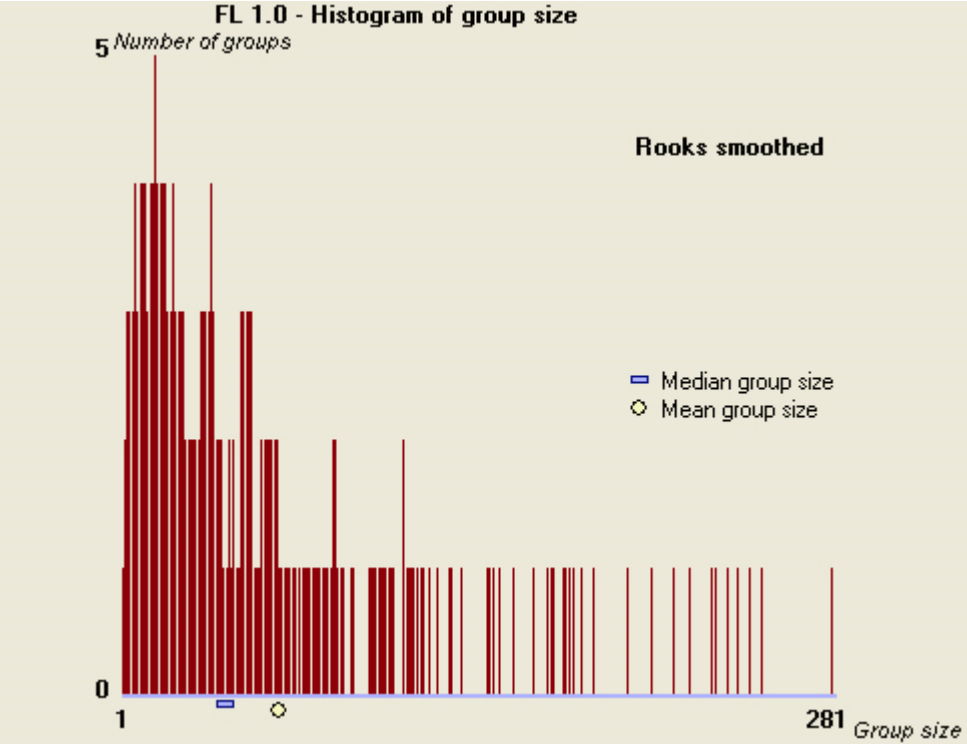
A group acts as a social environment of individuals: Great Woodswallows.

Many animals, including humans, tend to live in groups, herds, flocks, bands, packs, shoals, or colonies (hereafter: groups) of conspecific individuals. The size of these groups, as expressed by the number of participant individuals, is an important aspect of their social environment. Group size tend to be highly variable even within the same species, thus we often need statistical measures to quantify group size and statistical tests to compare these measures between two or more samples. Unfortunately, group size

measures are notoriously hard to handle statistically since groups size values typically exhibit an aggregated (right-skewed) distribution: most groups are small, few are large, and a very few are very large.

Statistical measures of group size roughly fall into two categories.

Outsiders' view of group size



Animal group size data tend to exhibit aggregated (right-skewed) distributions, i.e. most groups are small, a few are large, and a very few are very large. The distribution of rook colony sizes in Normandy, 1999-2000 (smoothed). Mean colony size is 60 pairs. (Data from Debout, 2003)

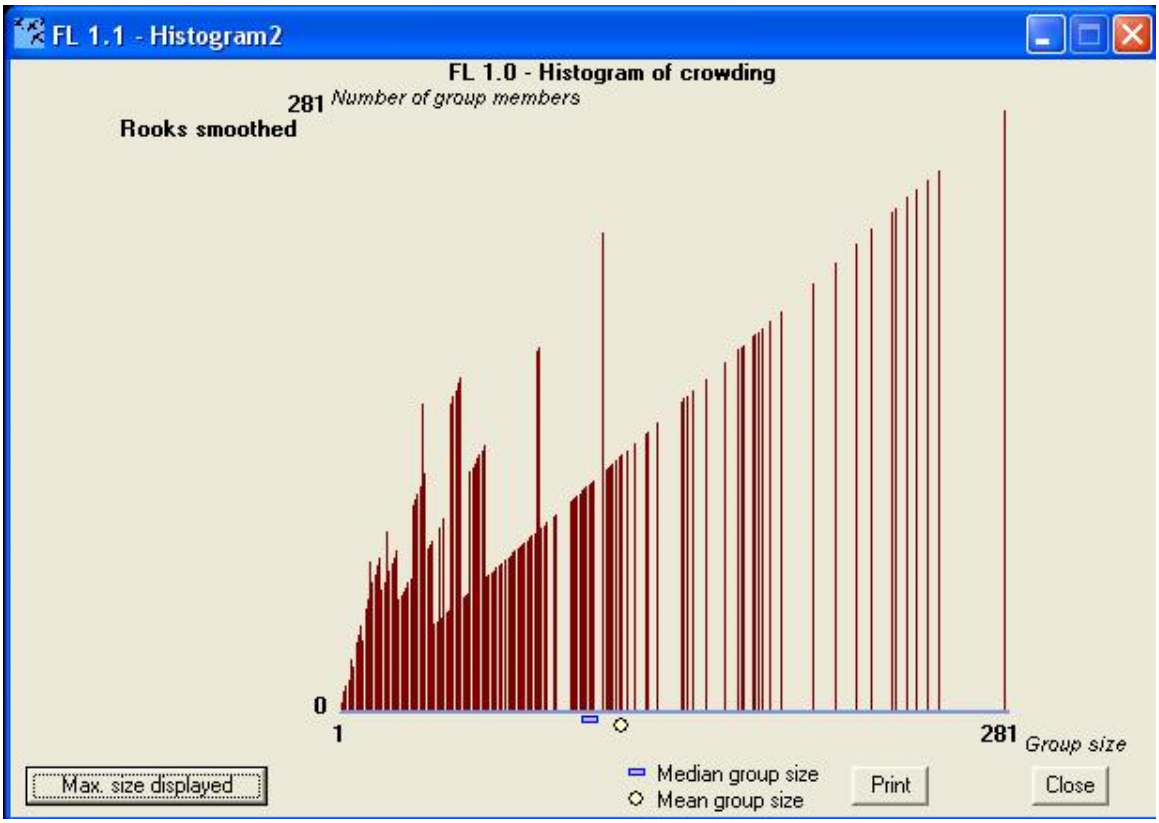
- Group size is the number of individuals within a group;
- **Mean group size** , i.e. the arithmetic mean of group sizes averaged across groups;
- **Confidence interval for mean group size**;
- **Median group size**, i.e. the median of group sizes calculated across groups;
- **Confidence interval for median group size**.

Insiders' view of group size

As Jarman (1974) pointed out, average individuals live in groups larger than average – simply because the groups smaller than average have fewer individuals than the groups larger than average. (Except for an unrealistic case when all groups are of equal size.) Therefore, when we wish to characterize a typical (average) individual's social

environment, we should not apply the outsiders' view of group size. Reiczigel et al. (2008) proposed the following measures:

- **Crowding** is the number of individuals within a group (equals to group size: 1 for a solitary individual, 2 for both individuals in a group of 2, etc.);
- **Mean crowding**, i.e. the arithmetic mean of crowding measures averaged across individuals (this was called "Typical Group Size" according to Jarman's 1974 terminology);
- **Confidence interval for mean crowding**.



Although large rook colonies are rare, however, they still incorporate a large proportion of individuals. Insiders' view of the same data set as above: the distribution of individuals (pairs) across colonies of different sizes. An average individual breeds in a colony of about 120 pairs, far larger than mean colony size.

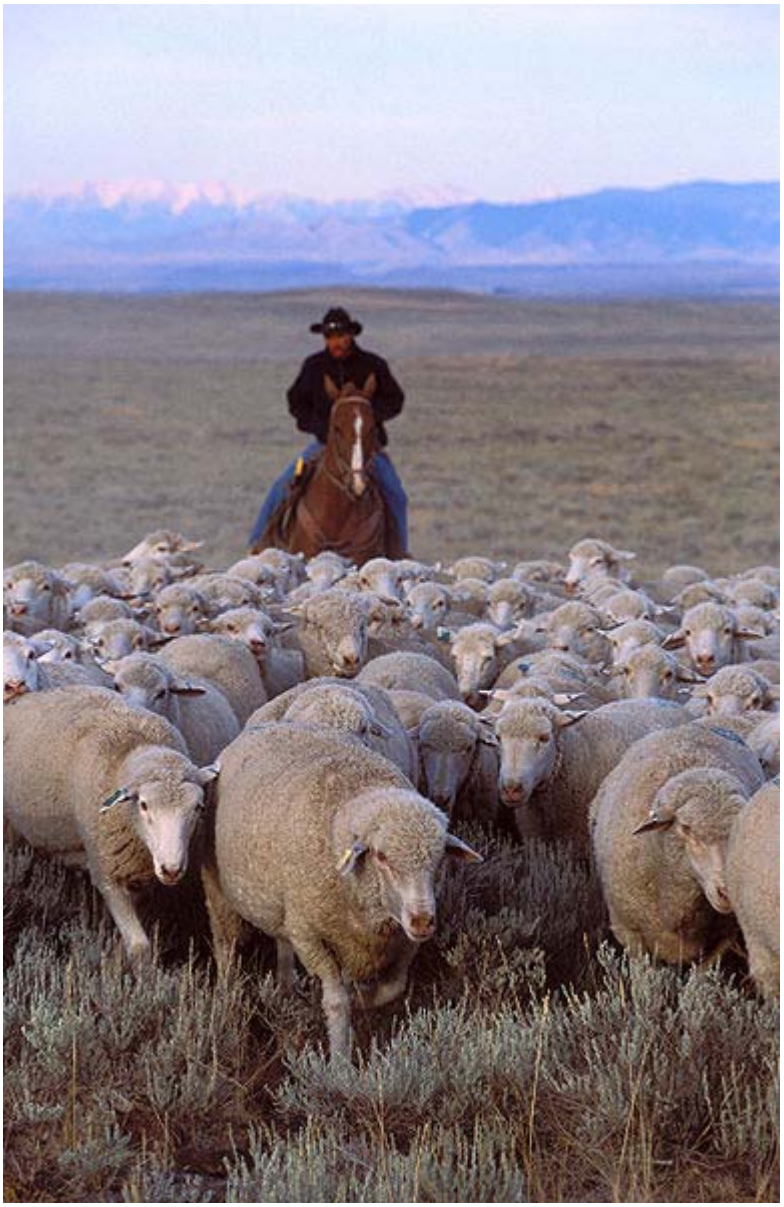
Statistical methods

Due to the aggregated (right-skewed) distribution of group members among groups, the application of parametric statistics would be misleading. Another problem arises when analyzing crowding values. Crowding data consist of nonindependent values, or ties, which show multiple and simultaneous changes due to a single biological event. (Say, all group members' crowding values change simultaneously whenever an individual joins or leaves.)

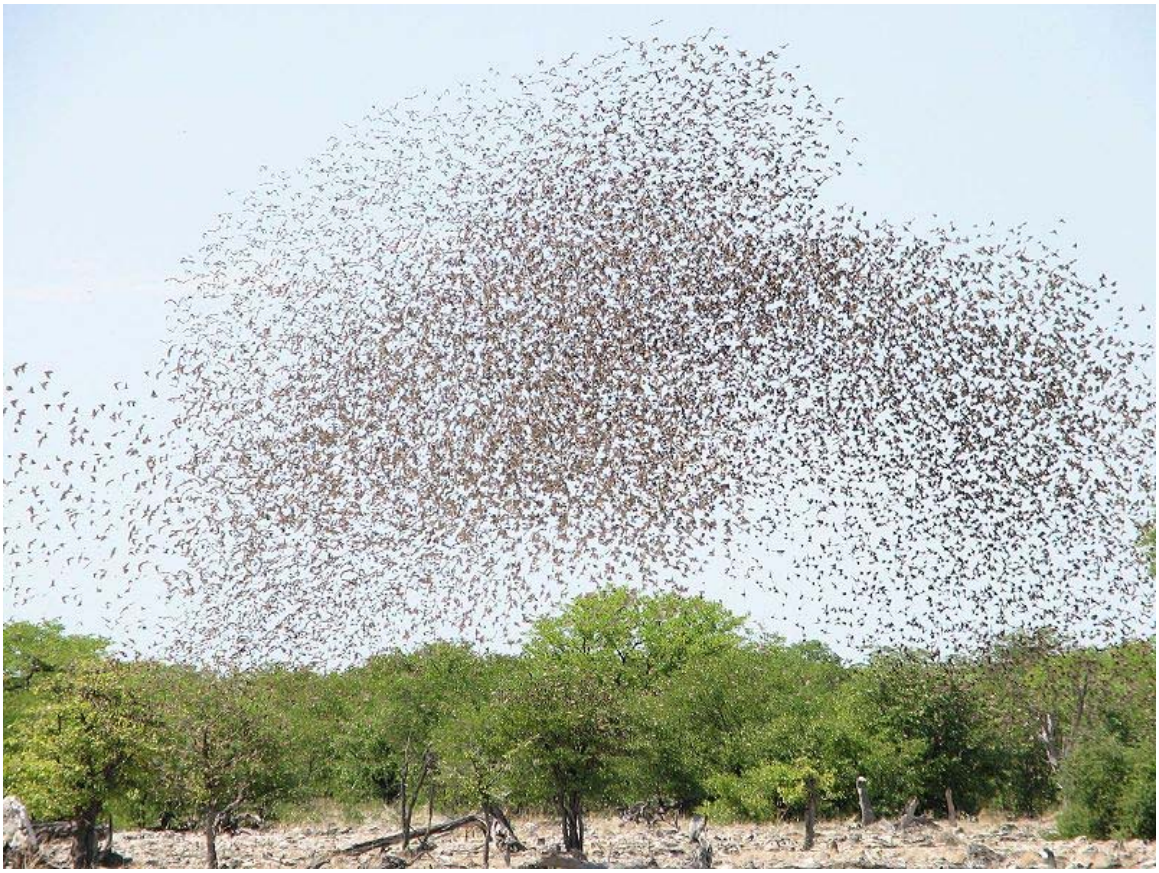
The paper by Reiczigel et al. (2008) discusses the statistical problems associated with group size measures (calculating confidence intervals, 2-sample tests, etc.) and offers a free statistical toolset (Flocker 1.1) to handle them in a user-friendly manner.

Gannet colony

African buffalo herd



Sheep flock



Red-billed Quelea flock

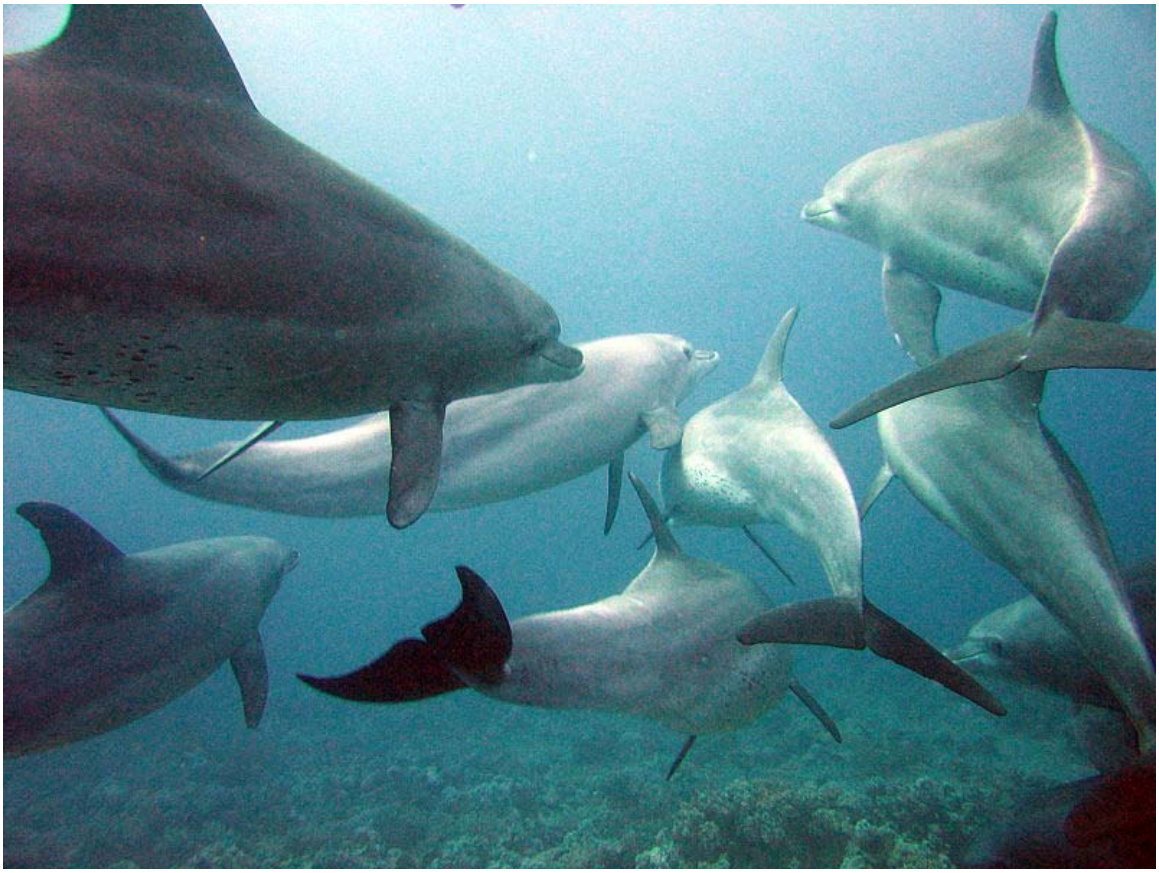
Bluestripe snapper schooling

Common Crane flock

Wolf pack

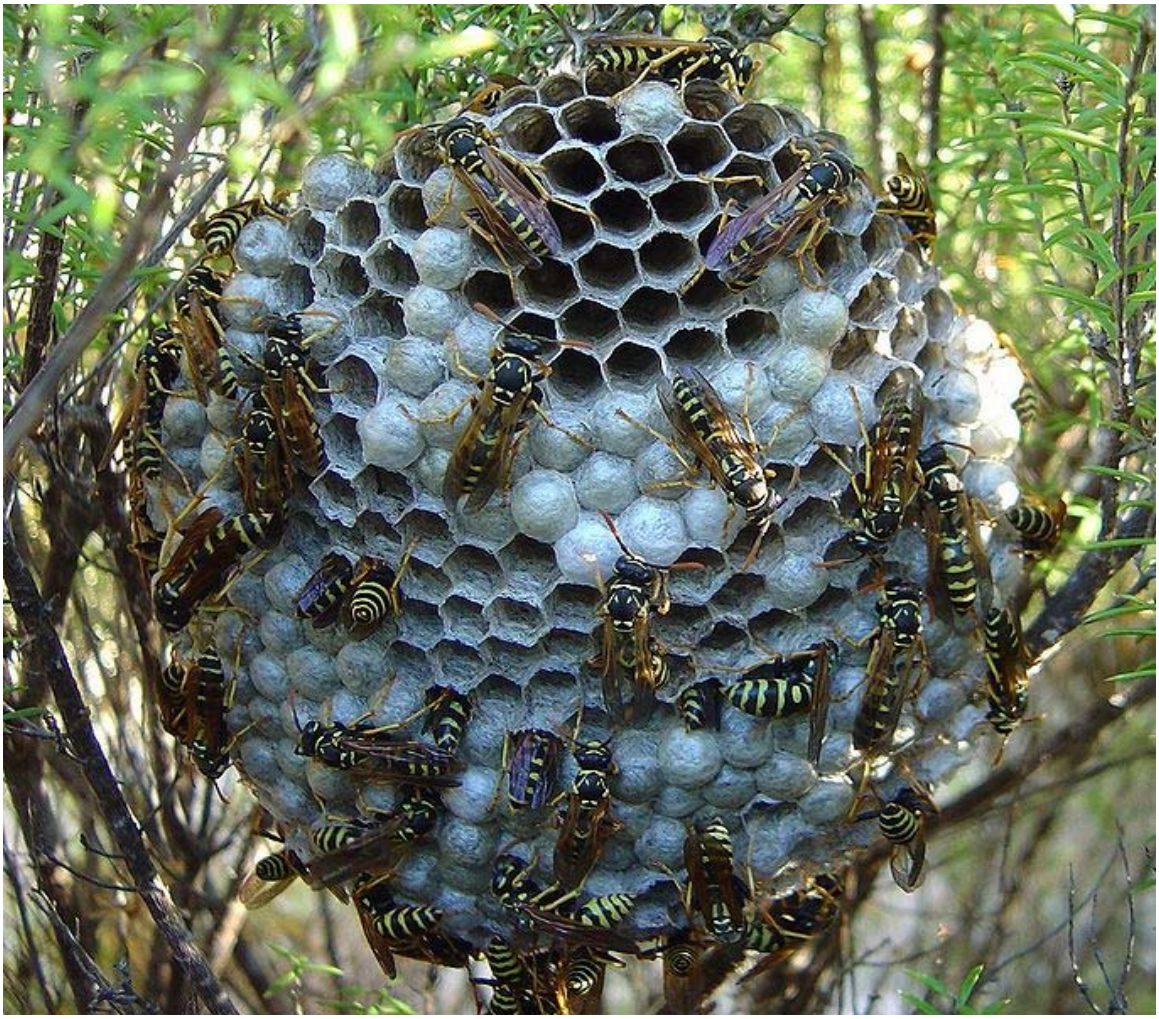
African Wild Dogs

Vicuñas



Bottlenose dolphins

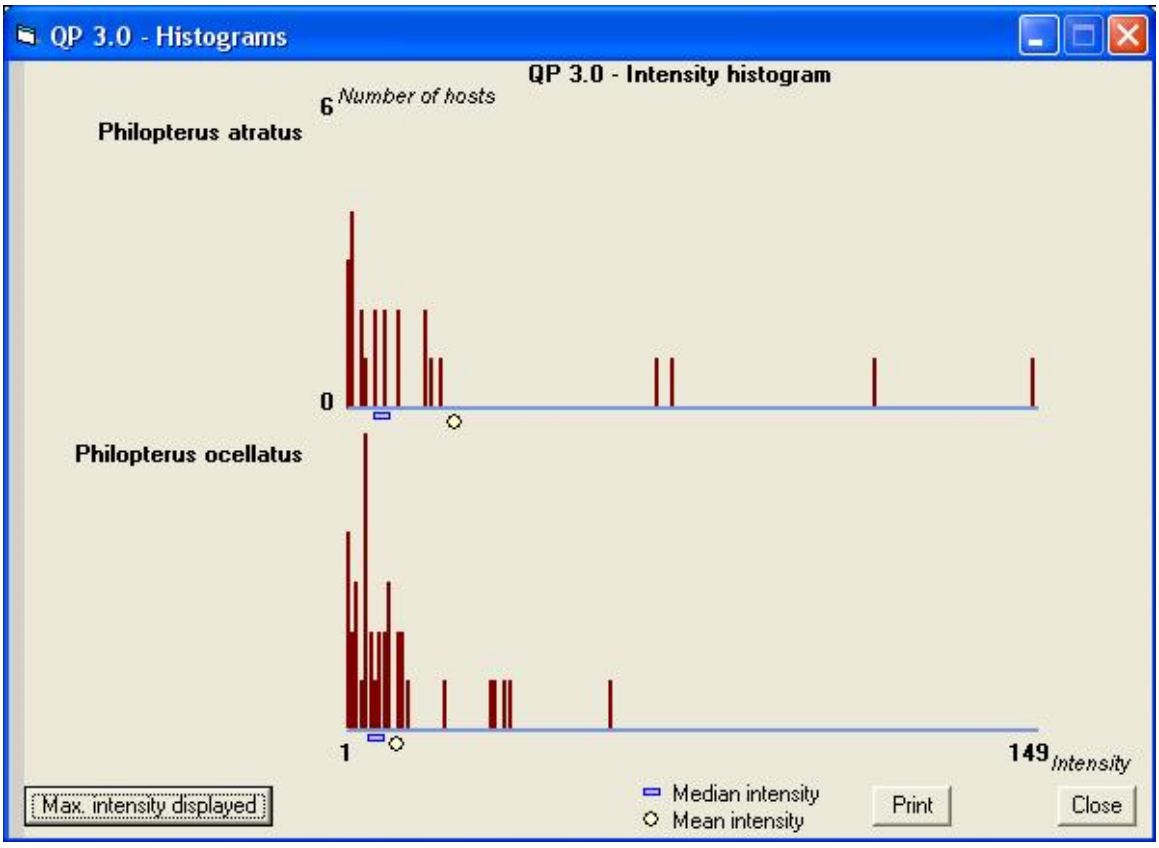
Flamingos



European Paper Wasp colony

Chapter- 5

Quantitative Parasitology



Intensity histograms are helpful to get a first impression about the differences of infection between 2 or more samples. Horizontal axis: infection classes, vertical axis: the number of host individuals belonging to each class.

Counting parasites

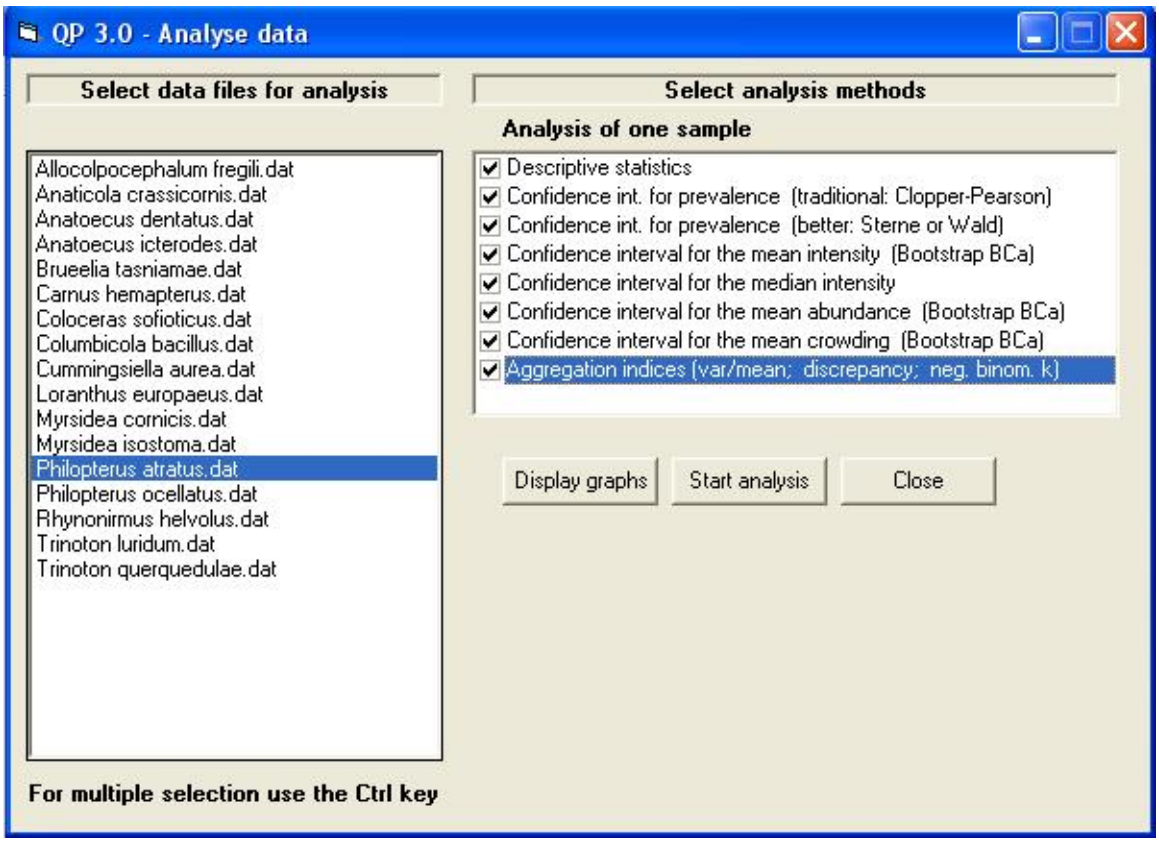
Quantifying parasites in a sample of hosts or comparing measures of infection across two or more samples can be challenging.

The parasitic infection of a sample of hosts inherently exhibits a complex pattern that cannot be adequately quantified by a single statistical measure. As the use of two or more separate indices is advisable, only two or more separate statistical tests can reliably compare infections different samples of hosts.

A few of the available statistical measures have markedly different biological interpretations, while others have more-or-less overlapping interpretations or no interpretations at all. Therefore, one should apply measures that have clear and separate biological interpretations thus do not predict each other.

Parasite individuals typically exhibit an aggregated (right-skewed) distribution among host individuals; most hosts harbour few if any parasites and a few hosts harbour many of them. This quantitative feature of parasitism renders many of traditional statistical methods obsolete and requires the use of advanced computer-intensive statistical methods.

How to describe the parasitic infection of a sample of hosts



Statistical procedures to characterize the infection/infestation of a sample of hosts.

Always give the host **sample size**. In most cases, this is expressed as the number of hosts individuals examined. (Exceptionally, other units may also be used for special cases.)

Describe **prevalence**. This is the proportion of infected hosts among all the hosts examined. Give the confidence interval (CI) of prevalence (either as a Clopper-Pearson interval or as adjusted Wald/Sterne's interval) to indicate the accuracy of the estimation (use of the confidence intervals belonging to the 95% probability is advisable).

Describe **mean intensity**. This is the mean number of parasites found in the infected hosts (the zeros of uninfected hosts are excluded). Since sample size and prevalence are known, mean intensity defines the quantity of parasites found in the sample of hosts. Given the typical aggregated (right-skewed) distribution of parasites, its actual value is highly dependent on a few extremely infected hosts. Also give CI to indicate the accuracy of the estimation. Use bias-corrected and accelerated bootstrap (BCa Bootsrap) to get this confidence interval.

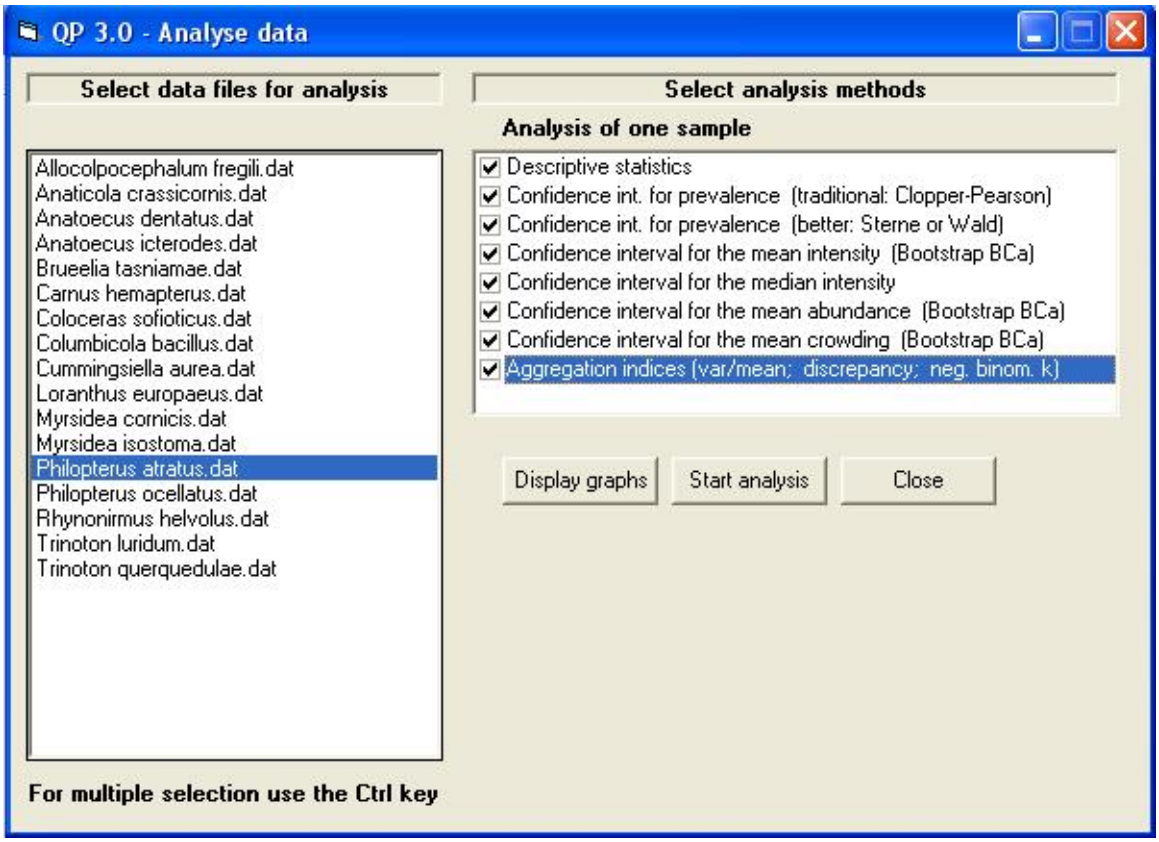
Describe **median intensity**. This is the median number of parasites found in infected hosts (the zeros of uninfected hosts are excluded). Median intensity shows a typical level of infection among the infected hosts. Use exact CI to indicate the accuracy of the estimation.

In certain cases one may prefer to use **mean abundance** instead of mean intensity. This is the mean number of parasites found in all hosts (involves the zero values of uninfected hosts). Give BCa Bootsrap confidence interval to indicate the accuracy of this estimation. This measure unifies two of the former ones: prevalence and mean intensity. Do not use it, unless you have a clearly specified a reason why to prefer it.

Describing **mean crowding** (intensity values averaged across parasite individuals) and its confidence interval is essential only for those who study density-dependent characters of parasites. BCa Bootsrap CI can be used to indicate the accuracy of the estimation.

Finally, quantify **levels of skewness** of the parasites' distribution among hosts. There are 3 indices widely used for this purpose, but their interpretation is quite similar. They predict each other rather well, thus it is not necessary to use all the 3 of them.

How to compare the parasite burdens across two or more samples



Statistical procedures to compare levels of infection/infestation across two or more samples of hosts

Compare **prevalences by Fisher's exact test**. This will show whether the proportion of infected individuals differs significantly between the two (or more) samples. The time need of this test may increase dramatically when several samples are involved. The use Chi-square test for the same purpose may be advisable in such cases.

Compare **mean intensities by a Bootstrap t-test**. This will show whether parasite quantities differ significantly between the infected proportions of the two samples.

Compare **median intensities by Mood's median test**. This will show whether the typical level of infection differs significantly between the infected proportions of the two samples.

One can also compare the **frequency distributions of intensities by a Stochastic equality test**. It compares several random pairs of individual values taken from the two samples to test whether or not there is a significant tendency to get higher values from one sample than from the other.

In certain cases, one may also decide to compare **mean abundances by a Bootstrap t-test**. This will show whether parasite quantities differ significantly between two samples. This comparison unifies two of the former ones: the comparison of prevalences and the comparison of mean intensities.

Finally, **mean crowding can be compared** across samples by a simple method: provided that the two 97.5% confidence intervals do not overlap, we conclude that the two values are different at a 95% level of significance.

Avoid typical mistakes

Do not use geometric mean because this measure is hard to interpret biologically.

Do not apply the usual form of *arithmetic mean ± standard deviation (mean ± SD)* to describe levels of infection because this is useful only for normal distributions, and not for the aggregated (right-skewed) distributions that characterize parasites. Use confidence intervals to quantify the accuracy of estimations.

Chapter- 6

Clinical Trial

Clinical trials are conducted to allow safety and efficacy data to be collected for health interventions (e.g., drugs, diagnostics, devices, therapy protocols). These trials can take place only after satisfactory information has been gathered on the quality of the non-clinical safety, and Health Authority/Ethics Committee approval is granted in the country where the trial is taking place.

Depending on the type of product and the stage of its development, investigators enroll healthy volunteers and/or patients into small pilot studies initially, followed by larger scale studies in patients that often compare the new product with the currently prescribed treatment. As positive safety and efficacy data are gathered, the number of patients is typically increased. Clinical trials can vary in size from a single center in one country to multicenter trials in multiple countries.

Due to the sizable cost a full series of clinical trials may incur, the burden of paying for all the necessary people and services is usually borne by the sponsor who may be a governmental organization, a pharmaceutical, or biotechnology company. Since the diversity of roles may exceed resources of the sponsor, often a clinical trial is managed by an outsourced partner such as a contract research organization or a clinical trials unit in the academic sector.

Overview

Clinical trials often involve patients with specific health conditions who then benefit from receiving otherwise unavailable treatments. In early phases, participants are healthy volunteers who receive financial incentives for their inconvenience. During dosing periods, study subjects typically remain on site at the unit for durations of anything from 1 to 30 nights, occasionally longer, although is not always required.

In planning a clinical trial, the sponsor or investigator first identifies the medication or device to be tested. Usually, one or more pilot experiments are conducted to gain insights for design of the clinical trial to follow. In medical jargon, effectiveness is how well a treatment works in practice and efficacy is how well it works in a clinical trial. In the

U.S., the elderly comprise only 14% of the population but they consume over one-third of drugs. Despite this, they are often excluded from trials because their more frequent health issues and drug use produces unreliable data. Women, children, and people with unrelated medical conditions are also frequently excluded.

In coordination with a panel of expert investigators (usually physicians well known for their publications and clinical experience), the sponsor decides what to compare the new agent with (one or more existing treatments or a placebo), and what kind of patients might benefit from the medication or device. If the sponsor cannot obtain enough patients with this specific disease or condition at one location, then investigators at other locations who can obtain the same kind of patients to receive the treatment would be recruited into the study.

During the clinical trial, the investigators: recruit patients with the predetermined characteristics, administer the treatment(s), and collect data on the patients' health for a defined time period. These patients are voluntaries and they are not paid for participating in clinical trials. These data include measurements like vital signs, concentration of the study drug in the blood, and whether the patient's health improves or not. The researchers send the data to the trial sponsor who then analyzes the pooled data using statistical tests.

Some examples of what a clinical trial may be designed to do:

- Assess the safety and effectiveness of a new medication or device on a specific kind of patient (e.g., patients who have been diagnosed with Alzheimer's disease)
- Assess the safety and effectiveness of a different dose of a medication than is commonly used (e.g., 10 mg dose instead of 5 mg dose)
- Assess the safety and effectiveness of an already marketed medication or device for a new indication, i.e. a disease for which the drug is not specifically approved
- Assess whether the new medication or device is more effective for the patient's condition than the already used, standard medication or device ("the gold standard" or "standard therapy")
- Compare the effectiveness in patients with a specific disease of two or more already approved or common interventions for that disease (e.g., Device A vs. Device B, Therapy A vs. Therapy B)

Note that while most clinical trials compare two medications or devices, some trials compare three or four medications, doses of medications, or devices against each other.

Except for very small trials limited to a single location, the clinical trial design and objectives are written into a document called a clinical trial protocol. The protocol is the 'operating manual' for the clinical trial, and ensures that researchers in different locations all perform the trial in the same way on patients with the same characteristics. (This uniformity is designed to allow the data to be pooled.) A protocol is always used in multicenter trials.

Because the clinical trial is designed to test hypotheses and rigorously monitor and assess what happens, clinical trials can be seen as the application of the scientific method to understanding human or animal biology.

Synonyms for 'clinical trials' include clinical studies, research protocols and clinical research.

The most commonly performed clinical trials evaluate new drugs, medical devices (like a new catheter), biologics, psychological therapies, or other interventions. Clinical trials may be required before the national regulatory authority approves marketing of the drug or device, or a new dose of the drug, for use on patients.

Beginning in the 1980s, harmonization of clinical trial protocols was shown as feasible across countries of the European Union. At the same time, coordination between Europe, Japan and the United States led to a joint regulatory-industry initiative on international harmonization named after 1990 as the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) Currently, most clinical trial programs follow ICH guidelines, aimed at "ensuring that good quality, safe and effective medicines are developed and registered in the most efficient and cost-effective manner. These activities are pursued in the interest of the consumer and public health, to prevent unnecessary duplication of clinical trials in humans and to minimize the use of animal testing without compromising the regulatory obligations of safety and effectiveness."

History

The history of clinical trials before 1750 is brief.

The concepts behind clinical trials, however, are ancient. The Book of Daniel verses 12 through 15, for instance, describes a planned experiment with both baseline and follow-up observations of two groups who either partook of, or did not partake of, "the King's meat" over a trial period of ten days. Persian physician and philosopher, Avicenna, gave such inquiries a more formal structure. In *The Canon of Medicine* in 1025 AD, he laid down rules for the experimental use and testing of drugs and wrote a precise guide for practical experimentation in the process of discovering and proving the effectiveness of medical drugs and substances. He laid out the following rules and principles for testing the effectiveness of new drugs and medications:

1. The drug must be free from any extraneous accidental quality.
2. It must be used on a simple, not a composite, disease.
3. The drug must be tested with two contrary types of diseases, because sometimes a drug cures one disease by its essential qualities and another by its accidental ones.
4. The quality of the drug must correspond to the strength of the disease. For example, there are some drugs whose heat is less than the coldness of certain diseases, so that they would have no effect on them.

5. The time of action must be observed, so that essence and accident are not confused.
6. The effect of the drug must be seen to occur constantly or in many cases, for if this did not happen, it was an accidental effect.
7. The experimentation must be done with the human body, for testing a drug on a lion or a horse might not prove anything about its effect on man.

One of the most famous clinical trials was James Lind's demonstration in 1747 that citrus fruits cure scurvy. He compared the effects of various different acidic substances, ranging from vinegar to cider, on groups of afflicted sailors, and found that the group who were given oranges and lemons had largely recovered from scurvy after 6 days.

Frederick Akbar Mahomed (d. 1884), who worked at Guy's Hospital in London, made substantial contributions to the process of clinical trials during his detailed clinical studies, where "he separated chronic nephritis with secondary hypertension from what we now term essential hypertension." He also founded "the Collective Investigation Record for the British Medical Association; this organization collected data from physicians practicing outside the hospital setting and was the precursor of modern collaborative clinical trials."

Types

One way of classifying clinical trials is by the way the researchers behave.

- In an observational study, the investigators observe the subjects and measure their outcomes. The researchers do not actively manage the study. An example is the Nurses' Health Study.
- In an interventional study, the investigators give the research subjects a particular medicine or other intervention. Usually, they compare the treated subjects to subjects who receive no treatment or standard treatment. Then the researchers measure how the subjects' health changes.

Another way of classifying trials is by their purpose. The U.S. National Institutes of Health (NIH) organizes trials into five (5) different types:

- *Prevention trials*: look for better ways to prevent disease in people who have never had the disease or to prevent a disease from returning. These approaches may include medicines, vitamins, vaccines, minerals, or lifestyle changes.
- *Screening trials*: test the best way to detect certain diseases or health conditions.
- *Diagnostic trials*: conducted to find better tests or procedures for diagnosing a particular disease or condition.
- *Treatment trials*: test experimental treatments, new combinations of drugs, or new approaches to surgery or radiation therapy.
- *Quality of life trials*: explore ways to improve comfort and the quality of life for individuals with a chronic illness (a.k.a. Supportive Care trials).

- *Compassionate use trials* or *expanded access*: provide partially tested, unapproved therapeutics prior to a small number of patients that have no other realistic options. Usually, this involves a disease for which no effective therapy exists, or a patient that has already attempted and failed all other standard treatments and whose health is so poor that he does not qualify for participation in randomized clinical trials. Usually, case by case approval must be granted by both the FDA and the pharmaceutical company for such exceptions.

Design

A fundamental distinction in evidence-based medicine is between observational studies and randomized controlled trials. Types of observational studies in epidemiology such as the cohort study and the case-control study provide less compelling evidence than the randomized controlled trial. In observational studies, the investigators only observe associations (correlations) between the treatments experienced by participants and their health status or diseases.

A randomized controlled trial is the study design that can provide the most compelling evidence that the study treatment causes the expected effect on human health.

Currently, some Phase II and most Phase III drug trials are designed as randomized, double blind, and placebo-controlled.

- Randomized: Each study subject is randomly assigned to receive either the study treatment or a placebo.
- Blind: The subjects involved in the study do not know which study treatment they receive. If the study is double-blind, the researchers also do not know which treatment is being given to any given subject. This 'blinding' is to prevent biases, since if a physician knew which patient was getting the study treatment and which patient was getting the placebo, he/she might be tempted to give the (presumably helpful) study drug to a patient who could more easily benefit from it. In addition, a physician might give extra care to only the patients who receive the placebos to compensate for their ineffectiveness. A form of double-blind study called a "double-dummy" design allows additional insurance against bias or placebo effect. In this kind of study, all patients are given both placebo and active doses in alternating periods of time during the study.
- Placebo-controlled: The use of a placebo (fake treatment) allows the researchers to isolate the effect of the study treatment.

Although the term "clinical trials" is most commonly associated with the large, randomized studies typical of Phase III, many clinical trials are small. They may be "sponsored" by single physicians or a small group of physicians, and are designed to test simple questions. In the field of rare diseases sometimes the number of patients might be the limiting factor for a clinical trial. Other clinical trials require large numbers of participants (who may be followed over long periods of time), and the trial sponsor is a

private company, a government health agency, or an academic research body such as a university.

Active comparator studies

Of note, during the last ten years or so it has become a common practice to conduct "active comparator" studies (also known as "active control" trials). In other words, when a treatment exists that is clearly better than doing nothing for the subject (*i.e.* giving them the placebo), the alternate treatment would be a standard-of-care therapy. The study would compare the 'test' treatment to standard-of-care therapy.

A growing trend in the pharmacology field involves the use of third-party contractors to obtain the required comparator compounds. Such third parties provide expertise in the logistics of obtaining, storing, and shipping the comparators. As an advantage to the manufacturer of the comparator compounds, a well-established comparator sourcing agency can alleviate the problem of parallel importing (importing a patented compound for sale in a country outside the patenting agency's sphere of influence).

Clinical trial protocol

A clinical trial protocol is a document used to gain confirmation of the trial design by a panel of experts and adherence by all study investigators, even if conducted in various countries.

The protocol describes the scientific rationale, objective(s), design, methodology, statistical considerations, and organization of the planned trial. Details of the trial are also provided in other documents referenced in the protocol such as an Investigator's Brochure.

The protocol contains a precise study plan for executing the clinical trial, not only to assure safety and health of the trial subjects, but also to provide an exact template for trial conduct by investigators at multiple locations (in a "multicenter" trial) to perform the study in exactly the same way. This harmonization allows data to be combined collectively as though all investigators (referred to as "sites") were working closely together. The protocol also gives the study administrators (often a contract research organization or CRO) as well as the site team of physicians, nurses and clinic administrators a common reference document for site responsibilities during the trial.

The format and content of clinical trial protocols sponsored by pharmaceutical, biotechnology or medical device companies in the United States, European Union, or Japan has been standardized to follow Good Clinical Practice guidance issued by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Regulatory authorities in Canada and Australia also follow ICH guidelines. Some journals, e.g. *Trials*, encourage trialists to publish their protocols in the journal.

Design features

Informed consent

An essential component of initiating a clinical trial is to recruit study subjects following procedures using a signed document called "informed consent".

Informed consent is a legally-defined process of a person being told about key facts involved in a clinical trial before deciding whether or not to participate. To fully describe participation to a candidate subject, the doctors and nurses involved in the trial explain the details of the study using terms the person will understand. Foreign language translation is provided if the participant's native language is not the same as the study protocol.

The research team provides an informed consent document that includes trial details, such as its purpose, duration, required procedures, risks, potential benefits and key contacts. The participant then decides whether or not to sign the document in agreement. Informed consent is not an immutable contract, as the participant can withdraw at any time without penalty.

Statistical power

In designing a clinical trial, a sponsor must decide on the target number of patients who will participate. The sponsor's goal usually is to obtain a statistically significant result showing a significant difference in outcome (e.g., improvement percentage in the treatment of psoriasis using hydrocortisone after 42 days). between the groups of patients who receive the study treatment and those who receive a placebo or a different treatment. The number of patients required to give a statistically significant result depends on the question the trial wants to answer. For example, to show the effectiveness of a new drug in a non-curable disease as metastatic kidney cancer requires many fewer patients than in a highly curable disease as seminoma if the drug is compared to a placebo.

The number of patients enrolled in a study has a large bearing on the ability of the study to reliably detect the size of the effect of the study intervention. This is described as the "power" of the trial. The larger the sample size or number of participants in the trial, the greater the statistical power.

However, in designing a clinical trial, this consideration must be balanced with the fact that more patients make for a more expensive trial. The power of a trial is not a single, unique value; it estimates the ability of a trial to detect a difference of a particular size (or larger) between the treated (tested drug/device) and control (placebo or standard treatment) groups. By example, a trial of a lipid-lowering drug versus placebo with 100 patients in each group might have a power of .90 to detect a difference between patients receiving study drug and patients receiving placebo of 10 mg/dL or more, but only have a power of .70 to detect a difference of 5 mg/dL.

Placebo groups

Merely giving a treatment can have nonspecific effects, and these are controlled for by the inclusion of a placebo group. Subjects in the treatment and placebo groups are assigned randomly and blinded as to which group they belong. Since researchers can behave differently to subjects given treatments or placebos, trials are also doubled-blinded so that the researchers do not know to which group a subject is assigned.

Assigning a person to a placebo group can pose an ethical problem if it violates his or her right to receive the best available treatment. The Declaration of Helsinki provides guidelines on this issue.

Phases

Clinical trials involving new drugs are commonly classified into four phases. Each phase of the drug approval process is treated as a separate clinical trial. The drug-development process will normally proceed through all four phases over many years. If the drug successfully passes through Phases I, II, and III, it will usually be approved by the national regulatory authority for use in the general population. Phase IV are 'post-approval' studies.

Before pharmaceutical companies start clinical trials on a drug, they conduct extensive pre-clinical studies.

Pre-clinical studies

It involves in vitro (test tube or cell culture) and in vivo (animal) experiments using wide-ranging doses of the study drug to obtain preliminary efficacy, toxicity and pharmacokinetic information. Such tests assist pharmaceutical companies to decide whether a drug candidate has scientific merit for further development as an investigational new drug.

Phase 0

Phase 0 is a recent designation for exploratory, first-in-human trials conducted in accordance with the United States Food and Drug Administration's (FDA) 2006 Guidance on Exploratory Investigational New Drug (IND) Studies. Phase 0 trials are also known as human microdosing studies and are designed to speed up the development of promising drugs or imaging agents by establishing very early on whether the drug or agent behaves in human subjects as was expected from preclinical studies. Distinctive features of Phase 0 trials include the administration of single subtherapeutic doses of the study drug to a small number of subjects (10 to 15) to gather preliminary data on the agent's pharmacokinetics (what the drug does to the body) and pharmacodynamics (what the body does to the drugs).

A Phase 0 study gives no data on safety or efficacy, being by definition a dose too low to cause any therapeutic effect. Drug development companies carry out Phase 0 studies to rank drug candidates in order to decide which has the best pharmacokinetic parameters in humans to take forward into further development. They enable go/no-go decisions to be based on relevant human models instead of relying on sometimes inconsistent animal data.

Questions have been raised by experts about whether Phase 0 trials are useful, ethically acceptable, feasible, speed up the drug development process or save money, and whether there is room for improvement.

Phase I

Phase I trials are the first stage of testing in human subjects. Normally, a small (20-100) group of healthy volunteers will be selected. This phase includes trials designed to assess the safety (pharmacovigilance), tolerability, pharmacokinetics, and pharmacodynamics of a drug. These trials are often conducted in an inpatient clinic, where the subject can be observed by full-time staff. The subject who receives the drug is usually observed until several half-lives of the drug have passed. Phase I trials also normally include dose-ranging, also called dose escalation, studies so that the appropriate dose for therapeutic use can be found. The tested range of doses will usually be a fraction of the dose that causes harm in animal testing. Phase I trials most often include healthy volunteers. However, there are some circumstances when real patients are used, such as patients who have terminal cancer or HIV and lack other treatment options. "The reason for conducting the trial is to discover the point at which a compound is too poisonous to administer." Volunteers are paid an inconvenience fee for their time spent in the volunteer centre. Pay ranges from a small amount of money for a short period of residence, to a larger amount of up to approx $6000 depending on length of participation.

There are different kinds of Phase I trials:

SAD
: Single Ascending Dose studies are those in which small groups of subjects are given a single dose of the drug while they are observed and tested for a period of time. If they do not exhibit any adverse side effects, and the pharmacokinetic data is roughly in line with predicted safe values, the dose is escalated, and a new group of subjects is then given a higher dose. This is continued until pre-calculated pharmacokinetic safety levels are reached, or intolerable side effects start showing up (at which point the drug is said to have reached the Maximum tolerated dose (MTD).

MAD
: Multiple Ascending Dose studies are conducted to better understand the pharmacokinetics & pharmacodynamics of multiple doses of the drug. In these studies, a group of patients receives multiple low doses of the drug, while samples (of blood, and other fluids) are collected at various time points and analyzed to

understand how the drug is processed within the body. The dose is subsequently escalated for further groups, up to a predetermined level.

Food effect
: A short trial designed to investigate any differences in absorption of the drug by the body, caused by eating before the drug is given. These studies are usually run as a crossover study, with volunteers being given two identical doses of the drug on different occasions; one while fasted, and one after being fed.

Phase II

Once the initial safety of the study drug has been confirmed in Phase I trials, Phase II trials are performed on larger groups (20-300) and are designed to assess how well the drug works, as well as to continue Phase I safety assessments in a larger group of volunteers and patients. When the development process for a new drug fails, this usually occurs during Phase II trials when the drug is discovered not to work as planned, or to have toxic effects.

Phase II studies are sometimes divided into Phase IIA and Phase IIB.

- Phase IIA is specifically designed to assess dosing requirements (how much drug should be given).
- Phase IIB is specifically designed to study efficacy (how well the drug works at the prescribed dose(s)).

Some trials combine Phase I and Phase II, and test both efficacy and toxicity.

Trial design
: Some Phase II trials are designed as case series, demonstrating a drug's safety and activity in a selected group of patients. Other Phase II trials are designed as randomized clinical trials, where some patients receive the drug/device and others receive placebo/standard treatment. Randomized Phase II trials have far fewer patients than randomized Phase III trials. :)

Phase III

Phase III studies are randomized controlled multicenter trials on large patient groups (300–3,000 or more depending upon the disease/medical condition studied) and are aimed at being the definitive assessment of how effective the drug is, in comparison with current 'gold standard' treatment. Because of their size and comparatively long duration, Phase III trials are the most expensive, time-consuming and difficult trials to design and run, especially in therapies for chronic medical conditions.

It is common practice that certain Phase III trials will continue while the regulatory submission is pending at the appropriate regulatory agency. This allows patients to continue to receive possibly lifesaving drugs until the drug can be obtained by purchase. Other reasons for performing trials at this stage include attempts by the sponsor at "label

expansion" (to show the drug works for additional types of patients/diseases beyond the original use for which the drug was approved for marketing), to obtain additional safety data, or to support marketing claims for the drug. Studies in this phase are by some companies categorised as "Phase IIIB studies."

While not required in all cases, it is typically expected that there be at least two successful Phase III trials, demonstrating a drug's safety and efficacy, in order to obtain approval from the appropriate regulatory agencies such as FDA (USA), or the EMA (European Union), for example.

Once a drug has proved satisfactory after Phase III trials, the trial results are usually combined into a large document containing a comprehensive description of the methods and results of human and animal studies, manufacturing procedures, formulation details, and shelf life. This collection of information makes up the "regulatory submission" that is provided for review to the appropriate regulatory authorities in different countries. They will review the submission, and, it is hoped, give the sponsor approval to market the drug.

Most drugs undergoing Phase III clinical trials can be marketed under FDA norms with proper recommendations and guidelines, but in case of any adverse effects being reported anywhere, the drugs need to be recalled immediately from the market. While most pharmaceutical companies refrain from this practice, it is not abnormal to see many drugs undergoing Phase III clinical trials in the market.

Phase IV

Phase IV trial is also known as **Post Marketing Surveillance Trial**. Phase IV trials involve the safety surveillance (pharmacovigilance) and ongoing technical support of a drug after it receives permission to be sold. Phase IV studies may be required by regulatory authorities or may be undertaken by the sponsoring company for competitive (finding a new market for the drug) or other reasons (for example, the drug may not have been tested for interactions with other drugs, or on certain population groups such as pregnant women, who are unlikely to subject themselves to trials). The safety surveillance is designed to detect any rare or long-term adverse effects over a much larger patient population and longer time period than was possible during the Phase I-III clinical trials. Harmful effects discovered by Phase IV trials may result in a drug being no longer sold, or restricted to certain uses: recent examples involve cerivastatin (brand names Baycol and Lipobay), troglitazone (Rezulin) and rofecoxib (Vioxx).

Length

Clinical trials are only a small part of the research that goes into developing a new treatment. Potential drugs, for example, first have to be discovered, purified, characterized, and tested in labs (in cell and animal studies) before ever undergoing clinical trials. In all, about 1,000 potential drugs are tested before just one reaches the point of being tested in a clinical trial. For example, a new cancer drug has, on average, 6

years of research behind it before it even makes it to clinical trials. But the major holdup in making new cancer drugs available is the time it takes to complete clinical trials themselves. On average, about 8 years pass from the time a cancer drug enters clinical trials until it receives approval from regulatory agencies for sale to the public. Drugs for other diseases have similar timelines.

Some reasons a clinical trial might last several years:

- For chronic conditions like cancer, it takes months, if not years, to see if a cancer treatment has an effect on a patient.
- For drugs that are not expected to have a strong effect (meaning a large number of patients must be recruited to observe *any* effect), recruiting enough patients to test the drug's effectiveness (i.e., getting statistical power) can take several years.
- Only certain people who have the target disease condition are eligible to take part in each clinical trial. Researchers who treat these particular patients must participate in the trial. Then they must identify the desirable patients and obtain consent from them or their families to take part in the trial.

The biggest barrier to completing studies is the shortage of people who take part. All drug and many device trials target a subset of the population, meaning not everyone can participate. Some drug trials require patients to have unusual combinations of disease characteristics. It is a challenge to find the appropriate patients and obtain their consent, especially when they may receive no direct benefit (because they are not paid, the study drug is not yet proven to work, or the patient may receive a placebo). In the case of cancer patients, fewer than 5% of adults with cancer will participate in drug trials. According to the Pharmaceutical Research and Manufacturers of America (PhRMA), about 400 cancer medicines were being tested in clinical trials in 2005. Not all of these will prove to be useful, but those that are may be delayed in getting approved because the number of participants is so low.

For clinical trials involving a seasonal indication (such as airborne allergies, Seasonal Affective Disorder, influenza, and others), the study can only be done during a limited part of the year (such as Spring for pollen allergies), when the drug can be tested. This can be an additional complication on the length of the study, yet proper planning and the use of trial sites in the southern as well as northern hemispheres allows for year-round trials can reduce the length of the studies.

Clinical trials that do not involve a new drug usually have a much shorter duration. (Exceptions are epidemiological studies like the Nurses' Health Study.)

Administration

Clinical trials designed by a local investigator and (in the U.S.) federally funded clinical trials are almost always administered by the researcher who designed the study and applied for the grant. Small-scale device studies may be administered by the sponsoring company. Phase III and Phase IV clinical trials of new drugs are usually administered by

a contract research organization (CRO) hired by the sponsoring company. (The sponsor provides the drug and medical oversight.) A CRO is a company that is contracted to perform all the administrative work on a clinical trial. It recruits participating researchers, trains them, provides them with supplies, coordinates study administration and data collection, sets up meetings, monitors the sites for compliance with the clinical protocol, and ensures that the sponsor receives 'clean' data from every site. Recently, site management organizations have also been hired to coordinate with the CRO to ensure rapid IRB/IEC approval and faster site initiation and patient recruitment.

At a participating site, one or more research assistants (often nurses) do most of the work in conducting the clinical trial. The research assistant's job can include some or all of the following: providing the local Institutional Review Board (IRB) with the documentation necessary to obtain its permission to conduct the study, assisting with study start-up, identifying eligible patients, obtaining consent from them or their families, administering study treatment(s), collecting and statistically analyzing data, maintaining and updating data files during followup, and communicating with the IRB, as well as the sponsor and CRO.

Ethical conduct

Clinical trials are closely supervised by appropriate regulatory authorities. All studies that involve a medical or therapeutic intervention on patients must be approved by a supervising ethics committee before permission is granted to run the trial. The local ethics committee has discretion on how it will supervise noninterventional studies (observational studies or those using already collected data). In the U.S., this body is called the Institutional Review Board (IRB). Most IRBs are located at the local investigator's hospital or institution, but some sponsors allow the use of a central (independent/for profit) IRB for investigators who work at smaller institutions.

To be ethical, researchers must obtain the full and informed consent of participating human subjects. (One of the IRB's main functions is ensuring that potential patients are adequately informed about the clinical trial.) If the patient is unable to consent for him/herself, researchers can seek consent from the patient's legally authorized representative. In California, the state has prioritized the individuals who can serve as the legally authorized representative.

In some U.S. locations, the local IRB must certify researchers and their staff before they can conduct clinical trials. They must understand the federal patient privacy (HIPAA) law and good clinical practice. International Conference of Harmonisation Guidelines for Good Clinical Practice (ICH GCP) is a set of standards used internationally for the conduct of clinical trials. The guidelines aim to ensure that the "rights, safety and well being of trial subjects are protected".

The notion of informed consent of participating human subjects exists in many countries all over the world, but its precise definition may still vary.

Informed consent is clearly a *necessary* condition for ethical conduct but does not *ensure* ethical conduct. The final objective is to serve the community of patients or future patients in a best-possible and most responsible way. However, it may be hard to turn this objective into a well-defined quantified objective function. In some cases this can be done, however, as for instance for questions of when to stop sequential treatments, and then quantified methods may play an important role.

Additional ethical concerns are present when conducting clinical trials on children (pediatrics).

Safety

Responsibility for the safety of the subjects in a clinical trial is shared between the sponsor, the local site investigators (if different from the sponsor), the various IRBs that supervise the study, and (in some cases, if the study involves a marketable drug or device) the regulatory agency for the country where the drug or device will be sold.

For safety reasons, many clinical trials of drugs are designed to exclude women of childbearing age, pregnant women, and/or women who become pregnant during the study. In some cases the male partners of these women are also excluded or required to take birth control measures.

Sponsor

- Throughout the clinical trial, the sponsor is responsible for accurately informing the local site investigators of the true historical safety record of the drug, device or other medical treatments to be tested, and of any potential interactions of the study treatment(s) with already approved medical treatments. This allows the local investigators to make an informed judgment on whether to participate in the study or not.
- The sponsor is responsible for monitoring the results of the study as they come in from the various sites, as the trial proceeds. In larger clinical trials, a sponsor will use the services of a Data Monitoring Committee (DMC, known in the U.S. as a Data Safety Monitoring Board). This is an independent group of clinicians and statisticians. The DMC meets periodically to review the unblinded data that the sponsor has received so far. The DMC has the power to recommend termination of the study based on their review, for example if the study treatment is causing more deaths than the standard treatment, or seems to be causing unexpected and study-related serious adverse events.
- The sponsor is responsible for collecting adverse event reports from all site investigators in the study, and for informing all the investigators of the sponsor's judgment as to whether these adverse events were related or not related to the study treatment. This is an area where sponsors can slant their judgment to favor the study treatment.
- The sponsor and the local site investigators are jointly responsible for writing a site-specific informed consent that accurately informs the potential subjects of the

true risks and potential benefits of participating in the study, while at the same time presenting the material as briefly as possible and in ordinary language. FDA regulations and ICH guidelines both require that “the information that is given to the subject or the representative shall be in language understandable to the subject or the representative." If the participant's native language is not English, the sponsor must translate the informed consent into the language of the participant.

Local site investigators

- A physician's first duty is to his/her patients, and if a physician investigator believes that the study treatment may be harming subjects in the study, the investigator can stop participating at any time. On the other hand, investigators often have a financial interest in recruiting subjects, and can act unethically in order to obtain and maintain their participation.
- The local investigators are responsible for conducting the study according to the study protocol, and supervising the study staff throughout the duration of the study.
- The local investigator or his/her study staff are responsible for ensuring that potential subjects in the study understand the risks and potential benefits of participating in the study; in other words, that they (or their legally authorized representatives) give truly informed consent.
- The local investigators are responsible for reviewing all adverse event reports sent by the sponsor. (These adverse event reports contain the opinion of both the investigator at the site where the adverse event occurred, and the sponsor, regarding the relationship of the adverse event to the study treatments). The local investigators are responsible for making an independent judgment of these reports, and promptly informing the local IRB of all serious and study-treatment-related adverse events.
- When a local investigator is the sponsor, there may not be formal adverse event reports, but study staff at all locations are responsible for informing the coordinating investigator of anything unexpected.
- The local investigator is responsible for being truthful to the local IRB in all communications relating to the study.

IRBs

Approval by an IRB, or ethics board, is necessary before all but the most informal medical research can begin.

- In commercial clinical trials, the study protocol is not approved by an IRB before the sponsor recruits sites to conduct the trial. However, the study protocol and procedures have been tailored to fit generic IRB submission requirements. In this case, and where there is no independent sponsor, each local site investigator submits the study protocol, the consent(s), the data collection forms, and supporting documentation to the local IRB. Universities and most hospitals have

in-house IRBs. Other researchers (such as in walk-in clinics) use independent IRBs.
- The IRB scrutinizes the study for both medical safety and protection of the patients involved in the study, before it allows the researcher to begin the study. It may require changes in study procedures or in the explanations given to the patient. A required yearly "continuing review" report from the investigator updates the IRB on the progress of the study and any new safety information related to the study.

Regulatory agencies

- If a clinical trial concerns a new regulated drug or medical device (or an existing drug for a new purpose), the appropriate regulatory agency for each country where the sponsor wishes to sell the drug or device is supposed to review all study data before allowing the drug/device to proceed to the next phase, or to be marketed. However, if the sponsor withholds negative data, or misrepresents data it has acquired from clinical trials, the regulatory agency may make the wrong decision.
- In the U.S., the FDA can audit the files of local site investigators after they have finished participating in a study, to see if they were correctly following study procedures. This audit may be random, or for cause (because the investigator is suspected of fraudulent data). Avoiding an audit is an incentive for investigators to follow study procedures.

Different countries have different regulatory requirements and enforcement abilities. "An estimated 40 percent of all clinical trials now take place in Asia, Eastern Europe, central and south America. “There is no compulsory registration system for clinical trials in these countries and many do not follow European directives in their operations”, says Dr. Jacob Sijtsma of the Netherlands-based WEMOS, an advocacy health organisation tracking clinical trials in developing countries."

Accidents

In March 2006 the drug TGN1412 caused catastrophic systemic organ failure in the individuals receiving the drug during its first human clinical trials (Phase I) in Great Britain. Following this, an Expert Group on Phase One Clinical Trials published a report. Investigational drug Trovan was tested on children in Nigeria causing severe health problems leading to lawsuits. In May 2010, a Phase III clinical trial for rheumatoid arthritis using ocrelizumab, an investigational new drug sponsored by Roche and Biogen Idec, was shut down after an excess number of deaths due to opportunistic infections in the interventional arm of the study. In October 2010, a Phase II trial for multiple sclerosis using the same drug was shut down after a patient died from systemic inflammatory response syndrome while taking the drug.

Economics

Sponsor

The cost of a study depends on many factors, especially the number of sites that are conducting the study, the number of patients required, and whether the study treatment is already approved for medical use. Clinical trials follow a standardized process.

The costs to a pharmaceutical company of administering a Phase III or IV clinical trial may include, among others:

- manufacturing the drug(s)/device(s) tested
- staff salaries for the designers and administrators of the trial
- payments to the contract research organization, the site management organization (if used) and any outside consultants
- payments to local researchers (and their staffs) for their time and effort in recruiting patients and collecting data for the sponsor
- study materials and shipping
- communication with the local researchers, including onsite monitoring by the CRO before and (in some cases) multiple times during the study
- one or more investigator training meetings
- costs incurred by the local researchers such as pharmacy fees, IRB fees and postage.
- any payments to patients enrolled in the trial (all payments are strictly overseen by the IRBs to ensure that patients do not feel coerced to take part in the trial by overly attractive payments)

These costs are incurred over several years.

In the U.S. there is a 50% tax credit for sponsors of certain clinical trials.

National health agencies such as the U.S. National Institutes of Health offer grants to investigators who design clinical trials that attempt to answer research questions that interest the agency. In these cases, the investigator who writes the grant and administers the study acts as the sponsor, and coordinates data collection from any other sites. These other sites may or may not be paid for participating in the study, depending on the amount of the grant and the amount of effort expected from them.

Clinical trials are traditionally expensive and difficult to undertake. Using internet resources can, in some cases, reduce the economic burden.

Investigators

Many clinical trials do not involve any money. However, when the sponsor is a private company or a national health agency, investigators are almost always paid to participate. These amounts can be small, just covering a partial salary for research assistants and the

cost of any supplies (usually the case with national health agency studies), or be substantial and include 'overhead' that allows the investigator to pay the research staff during times in between clinical trials.

Patients

In Phase I drug trials, participants are paid because they give up their time (sometimes away from their homes) and are exposed to unknown risks, without the expectation of any benefit. In most other trials, however, patients are not paid, in order to ensure that their motivation for participating is the hope of getting better or contributing to medical knowledge, without their judgment being skewed by financial considerations. However, they are often given small payments for study-related expenses like travel or as compensation for their time in providing follow-up information about their health after they are discharged from medical care.

Participating in a clinical trial

Newspaper advertisements seeking patients and healthy volunteers to participate in clinical trials

Phase 0 and Phase I drug trials seek healthy volunteers. Most other clinical trials seek patients who have a specific disease or medical condition.

Locating trials

Depending on the kind of participants required, sponsors of clinical trials use various recruitment strategies, including patient databases, newspaper and radio advertisements, flyers, posters in places the patients might go (such as doctor's offices), and personal recruitment of patients by investigators.

Volunteers with specific conditions or diseases have additional online resources to help them locate clinical trials. For example, people with Parkinson's disease can use PDtrials to find up-to-date information on Parkinson's disease trials currently enrolling participants in the U.S. and Canada, and search for specific Parkinson's clinical trials using criteria such as location, trial type, and symptom. Other disease-specific services exist for volunteers to find trials related to their condition. Volunteers may also search directly on ClinicalTrials.gov to locate trials using a registry run by the U.S. National Institutes of Health and National Library of Medicine.

However, many clinical trials will not accept participants who contact them directly to volunteer as it is believed this may bias the characteristics of the population being studied. Such trials typically recruit via networks of medical professionals who ask their individual patients to consider enrollment.

Steps for volunteers

Before participating in a clinical trial, interested volunteers should speak with their doctors, family members, and others who have participated in trials in the past. After locating a trial, volunteers will often have the opportunity to speak or e-mail the clinical trial coordinator for more information and to answer any questions. After receiving consent from their doctors, volunteers then arrange an appointment for a screening visit with the trial coordinator.

All volunteers being considered for a trial are required to undertake a medical screen. There are different requirements for different trials, but typically volunteers will have the following tests:

- Measurement of the electrical activity of the heart (ECG)
- Measurement of blood pressure, heart rate and temperature
- Blood sampling
- Urine sampling
- Weight and height measurement
- Drugs abuse testing
- Pregnancy testing (females only)

Information technology

The last decade has seen a proliferation of information technology use in the planning and conduct of clinical trials. Clinical trial management systems (CTMS) are often used

by research sponsors or CROs to help plan and manage the operational aspects of a clinical trial, particularly with respect to investigational sites. Web-based electronic data capture (EDC) and clinical data management systems (CDMS) are used in a majority of clinical trials to collect case report data from sites, manage its quality and prepare it for analysis. Interactive voice response systems (IVRS) are used by sites to register the enrollment of patients using a phone and to allocate patients to a particular treatment arm (although phones are being increasingly replaced with web-based tools which are sometimes part of the EDC system). Patient-reported outcome measures are being increasingly collected using hand-held, sometimes wireless ePRO (or eDiary) devices. Statistical software is used to analyze the collected data and prepare it for regulatory submission. Access to many of these applications are increasingly aggregated in web-based clinical trial portals.

Criticism

Marcia Angell has been a stern critic of U.S. health care in general and the pharmaceutical industry in particular. She is scathing on the topic of how clinical trials are conducted in America:

Many drugs that are assumed to be effective are probably little better than placebos, but there is no way to know because negative results are hidden.... Because favorable results were published and unfavorable results buried ... the public and the medical profession believed these drugs were potent.... Clinical trials are also biased through designs for research that are chosen to yield favorable results for sponsors. For example, the sponsor's drug may be compared with another drug administered at a dose so low that the sponsor's drug looks more powerful. Or a drug that is likely to be used by older people will be tested in young people, so that side effects are less likely to emerge. A common form of bias stems from the standard practice of comparing a new drug with a placebo, when the relevant question is how it compares with an existing drug. In short, it is often possible to make clinical trials come out pretty much any way you want, which is why it's so important that investigators be truly disinterested in the outcome of their work.... It is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgment of trusted physicians or authoritative medical guidelines. I take no pleasure in this conclusion, which I reached slowly and reluctantly over my two decades as an editor of the *New England Journal of Medicine*.

Angell believes that members of medical school faculties who conduct clinical trials should not accept any payments from drug companies except research support, and that support should have no strings attached, including control by the companies over the design, interpretation, and publication of research results. She has speculated that "perhaps most" of the clinical trials are viewed by critics as "excuses to pay doctors to put patients on a company's already-approved drug".

Chapter- 7

Statistical Parametric Mapping and Clinical Utility of Diagnostic Tests

Statistical Parametric Mapping

Statistical parametric mapping or **SPM** is a statistical technique for examining differences in brain activity recorded during functional neuroimaging experiments using neuroimaging technologies such as fMRI or PET. It may also refer to a specific piece of software created by the *Wellcome Department of Imaging Neuroscience* (part of University College London) to carry out such analyses.

The statistical parametric mapping approach

Unit of measurement

Functional neuroimaging, one type of 'brain scanning', involves the measurement of brain activity. The specific technique used to measure brain activity depends on the imaging technology being used. Regardless of which technology is used, the scanner produces a 'map' of the area being scanned that is represented as voxels. Each voxel typically represents the activity of a particular coordinate in three dimensional space. The exact size of a voxel will vary depending on the technology used, although fMRI voxels typically represent a volume of 27 mm^3 (a cube with 3mm length sides).

Experimental design

Researchers are often interested in examining brain activity linked to a specific psychological process or processes. An experimental approach to this problem might involve asking the question 'which areas of the brain are significantly more active when a person is doing task A compared to task B?'. Although each task might be designed to be identical, except for the aspect of behaviour under investigation, the brain is still likely to show changes in activity between tasks due to factors other than task differences (as the brain is involved with co-ordinating a whole range of parallel functions unrelated to the

experimental task). Furthermore, the signal may contain noise from the imaging process itself.

To accommodate these random effects, and to highlight the areas of activity linked specifically to the process under investigation, statistics are used to look for the most significant difference above and beyond background brain activity. This involves a multi-stage process to prepare the data, and to subsequently analyse it using a statistical method known as the general linear model.

Image pre-processing

Images from the brain scanner may be pre-processed before any statistical comparison takes place to remove noise or correct for sampling errors.

A study will usually scan a subject several times. To account for the motion of the head between scans, the images will usually be adjusted so each of the voxels in the images corresponds (approximately) to the same site in the brain.

Functional neuroimaging studies usually involve several participants, who will have slightly differently shaped brains. All are likely to have the same gross anatomy, but there will be minor differences in overall brain size, individual variation in topography of the gyri and sulci of the cerebral cortex, and morphological differences in deep structures such as the corpus callosum. To aid comparisons, the 3D image of each brain is transformed so that superficial structures line up, a process known as *spatial normalization*. Such normalization typically involves not only translation and rotation, but also scaling and nonlinear warping of the brain surface to match a standard template. Standard brain maps such as the Talairach-Tournoux or templates from the Montréal Neurological Institute (MNI) are often used to allow researchers from across the world to compare their results.

Images are often smoothed (similar to the 'blur' effect used in some image-editing software) by which voxels are averaged with their neighbours, typically using a Gaussian filter or by wavelet transformation, to make the data less noisy.

Statistical comparison

Parametric statistical models are assumed at each voxel, using the general linear model to describe the variability in the data in terms of experimental and confounding effects, and residual variability. Hypotheses expressed in terms of the model parameters are assessed at each voxel with univariate statistics.

Analyses may also be conducted to examine differences over a time series (i.e correlations between a task variable and brain activity in a certain area) using linear convolution models of how the measured signal is caused by underlying changes in neural activity.

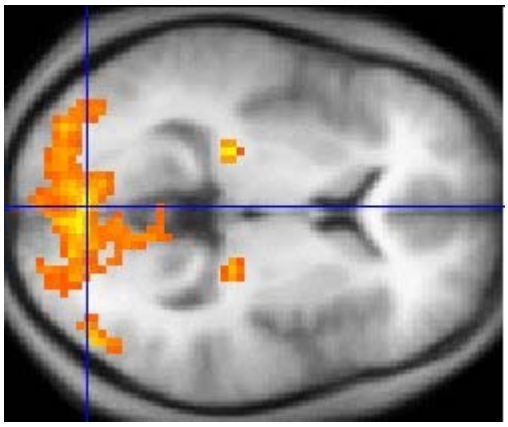
Because many statistical tests are being conducted, adjustments have to be made to control for Type I errors (false positives) potentially caused by the comparison of levels of activity at a large number of voxels. In this case, a Type I error would result in falsely detecting background brain activity as activity related to the task. Adjustments are made, based on the number of resels in the image and the theory of continuous random fields in order to set a new criterion for statistical significance that adjusts for the problem of multiple comparisons.

Graphical representations

Differences in measured brain activity can be represented in a number of ways.

Most simply, they can be presented as a table, displaying coordinates that show the most significant differences in activity between tasks. However, differences in brain activity are more often shown as patches of colour on an MRI brain 'slice', with the colours representing the location of voxels that have shown statistically significant differences between conditions. The gradient of color is mapped to statistical values, such as t-values or z-scores. This creates an intuitive and visually appealing means of delineating the relative statistical strength of a given area of activation. Recently, an alternative approach has been suggested, in which the statistical map is combined with the map of the original difference in brain activity (or, more generally speaking, with the original *contrast*) and colorcodes are attributed to the latter.

Differences in activity may also be represented as a 'glass brain', a representation of three outline views of the brain as if it were transparent. Only the patches of activation are visible as areas of shading. This is useful as a quick means of summarizing the total area of significant change in a given statistical comparison.



Brain activation from fMRI shown as patch of colour on MRI scan

SPM software

SPM is software written by the Wellcome Department of Imaging Neuroscience at University College London to aid in the analysis of functional neuroimaging data. It is written using MATLAB and is distributed as free software.

Clinical utility of diagnostic tests

The **clinical utility of a diagnostic test** is its capacity to rule diagnosis in and/or out and to make a decision possible to adopt or to reject a therapeutic action. It can be integrated into clinical prediction rules for specific diseases or outcomes.

Factors determining the utility

A. Association between test results and disease is a must.

B. The pre-test probability of a disease

C. The demand of the testee in regard to the post-test probability of disease according with the given test result in order to rule in or out the disease and to accept or reject a particular therapeutic action.

More explanation of the factors mentioned above

In regard to A. In the case of no association the post-test probability of disease is independent of the positivity or negativity of the test result and is always equal to the pre-test probability. In other words, the test result does not change the degree of (un)certainty of presence or absence of the target disease: the test is useless. If association occurs the degree of post test probability of disease increases with positive test results and decreases with negative ones. Association increases the degree of certainty by which the hypothesis of the disease can be adopted given a positive test result and by which the hypothesis can be rejected given a negative test result.

In regard to B. The pre-test probability of disease influences the post-test probability. Pre-test probability is the probability that a person suffers from a disease before the test is executed. A high pre-test probability will tend to allow (much) easier to confirm the hypothesis of the presence of the target disease, a low pre-test probability of disease will tend to allow (much) easier to accept the hypothesis of absence of the disease.

In regard to C. It is the user of the test (or the testee) who determines which degree of certainty that is needed to decide to the presence or absence of the target disease and/or to take the adeaquate decisions in regard to therapy. Thus it is possible that someone is of opinion that the test result is useful while another thinks that the test (on its own or in the given combination of other test results and/or data) is useless.

Conclusion

Out of B and C an important conclusion can be deducted. It is not because the association between a test result and the presence or absence of the disease is weak that the test is necessarily useless since a high pre-test probability and/or a 'low' degree of demanded certainty by tester and/or testee can make a test with a weak relation (a modest likelihood ratio) to the target disease useful (allow to decide to the presence or absence of the target disease given a positive test result despite the weak relationship). An analogue reasoning can be made for ruling out the target disease.

On the other hand a test result on a test showing a strong association with the absence or presence of the target disease can be useless because of a 'low' pre-test probability and/or a too high demand for the degree of certainty. An analogue reasoning can be made for ruling out the diagnosis of the target disease.

Every test that shows an association between test results and the target disease is potentially useful. If it is not on its own thought to be useful then combination of it with other test results and/or data can potentially lead to a post-test probability that is thought to be high enough to rule the diagnosis in or low enough to rule the diagnosis out.

Tests can be useful to rule disease in or out or to rule both disease in (positive test result) and out (negative test result)

An example

A formula for the calculation of the post-test probability of disease is given by:

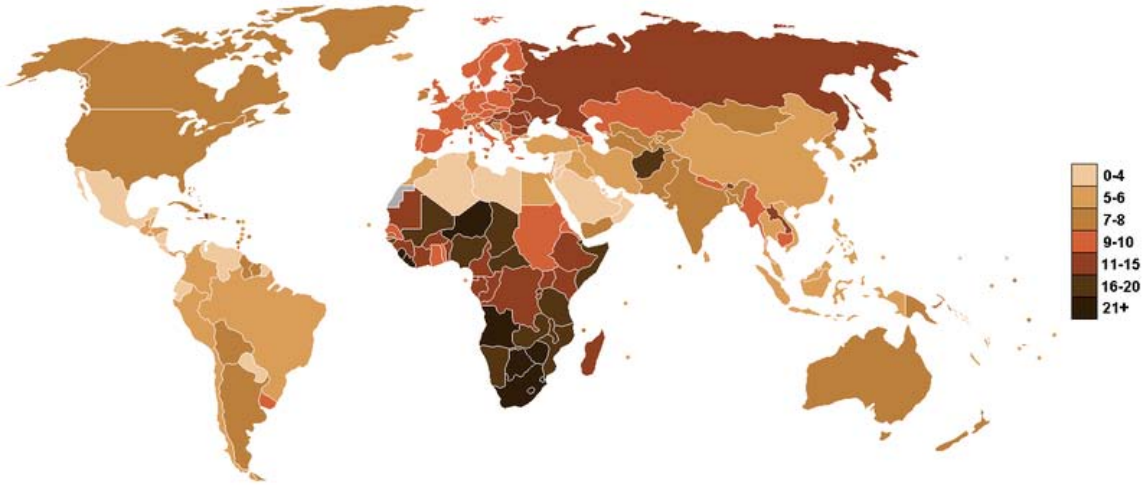
```
NK = PR*LR/(PR*(LR-1)-1)
```

Wherein NK = post-test probability of disease and PR = pre-test probability of disease and LR = likelihood ratio.

Let PR = .1 and LR+ = 10 and our demand for certainty = 95% then the post-test probability equals 52.6% and this is far insufficient. to accept the hypothesis of the presence of the target disease. Otherwise let PR = 90% and LR+ = 3 then NK = 96.4% what suffices to accept the presence of the target disease since the sufficient thougt degree of certainty was 95%. Although a LR+ = 10 points to a much greater association between a positive test result and the presence of the target disease than a LR+ = 3, an LR+ = 3 can suffice for ruling a disease in while it is possible that a LR+ = 10 does not suffice. If the demanded degree of certainty should have been as high as 97% then both pre-test probabilities and LR's should not have been sufficient to rule the diagnosis in. In this example the crucial role of the pre-test probability and the demand of the degree of certainty for the usefulness of a positive test result is illustrated.

Chapter- 8

Mortality Rate



Crude death rate by country

Mortality rate is a measure of the number of deaths (in general, or due to a specific cause) in some population, scaled to the size of that population, per unit time. Mortality rate is typically expressed in units of deaths per 1000 individuals per year; thus, a mortality rate of 9.5 in a population of 100,000 would mean 950 deaths per year in that entire population, or 0.95% out of the total. It is distinct from morbidity rate, which refers to the number of individuals in poor health during a given time period (the prevalence rate) or the number of newly appearing cases of the disease per unit of time (incidence rate).

One distinguishes:

1. The **crude death rate**, the total number of deaths per year per 1000 people. As of July 2009 the crude death rate for the whole world is about 8.37 per 1000 per year according to the current CIA World Factbook.
2. The **perinatal mortality rate**, the sum of neonatal deaths and fetal deaths (stillbirths) per 1000 births.

3. The **maternal mortality rate**, the number of maternal deaths per 100,000 women of reproductive age in same time period.
4. The **infant mortality rate**, the number of deaths of children less than 1 year old per 1000 live births.
5. The **child mortality rate**, the number of deaths of children less than 5 years old per 1000 live births.
6. The **standardised mortality rate** (SMR)- This represents a proportional comparison to the numbers of deaths that would have been expected if the population had been of a standard composition in terms of age, gender, etc.
7. The **age-specific mortality rate** (ASMR) - This refers to the total number of deaths per year per 1000 people of a given age (e.g. age 62 last birthday).

In regard to the success or failure of medical treatment or procedures, one would also distinguish:

1. The **early mortality rate**, the total number of deaths in the early stages of an ongoing treatment, or in the period immediately following an acute treatment.
2. The **late mortality rate**, the total number of deaths in the late stages of an ongoing treatment, or a significant length of time after an acute treatment.

Note that the crude death rate as defined above and applied to a whole population can give a misleading impression. The crude death rate depends on the age (and gender) specific mortality rates and the age (and gender) distribution of the population. The number of deaths per 1000 people can be higher for developed nations than in less-developed countries, despite life expectancy being higher in developed countries due to standards of health being better. This happens because developed countries typically have a completely different population age distribution, with a much higher proportion of older people, due to both lower recent birth rates and lower mortality rates. A more complete picture of mortality is given by a life table which shows the mortality rate separately for each age. A life table is necessary to give a good estimate of life expectancy.

Statistics

World historical and predicted crude death rates (1950-2050)
UN, medium variant, 2008 rev.

Years	CDR	Years	CDR
1950-1955	19.5	2000-2005	8.6
1955-1960	17.3	2005-2010	8.5
1960-1965	15.5	2010-2015	8.3
1965-1970	13.2	2015-2020	8.3

1970-1975	11.4	2020-2025	8.3
1975-1980	10.7	2025-2030	8.5
1980-1985	10.3	2030-2035	8.8
1985-1990	9.7	2035-2040	9.2
1990-1995	9.4	2040-2045	9.6
1995-2000	8.9	2045-2050	10

During ancient times and the Middle Ages, the crude death rate was about 40 deaths per year per 1,000 people.

The ten countries with the highest crude death rate, according to the 2009 CIA World Factbook estimates, are:

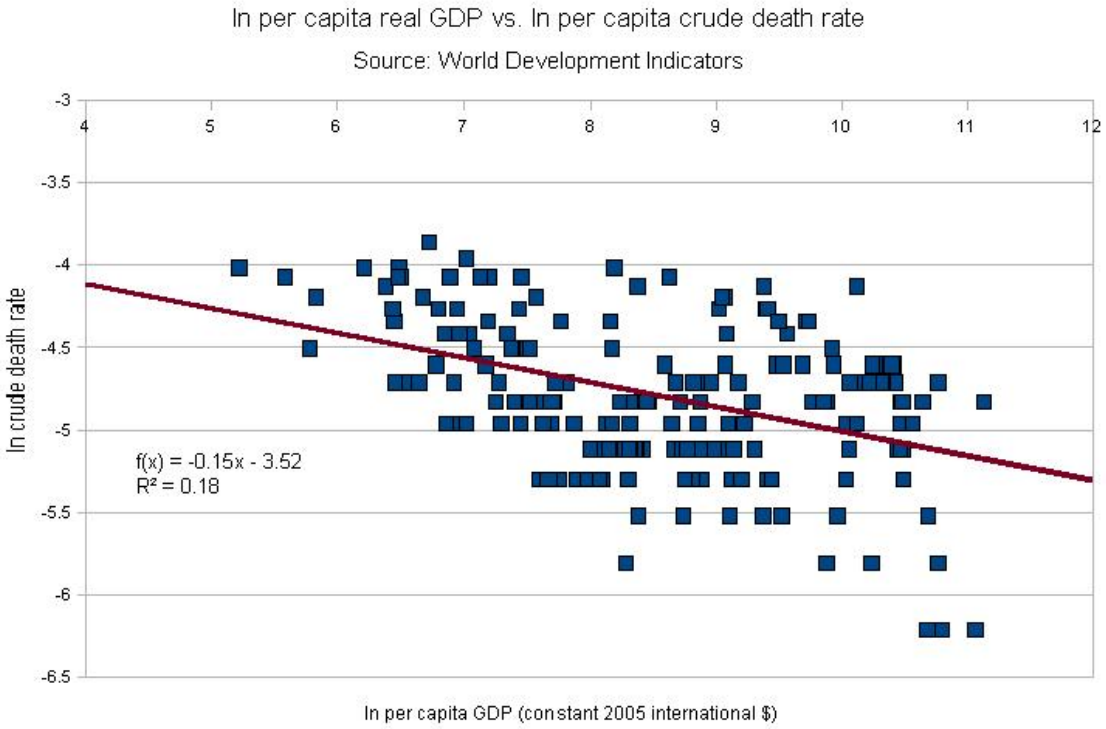
Rank	Country	Death rate (deaths/1000 persons)
1	Swaziland	30.83
2	Angola	24.08
3	Lesotho	22.20
4	Sierra Leone	21.91
5	Zambia	21.34
6	Liberia	20.73
7	Mozambique	20.07
8	Afghanistan	19.18
9	Djibouti	19.10
10	Central African Republic	17.84

According to the World Health Organization, the 10 leading causes of death in 2002 were:

1. 12.6% Ischaemic heart disease
2. 9.7% Cerebrovascular disease
3. 6.8% Lower respiratory infections
4. 4.9% HIV/AIDS
5. 4.8% Chronic obstructive pulmonary disease
6. 3.2% Diarrhoeal diseases
7. 2.7% Tuberculosis
8. 2.2% Trachea/bronchus/lung cancers

9. 2.2% Malaria
10. 2.1% Road traffic accidents

Causes of death vary greatly between first and third world countries.



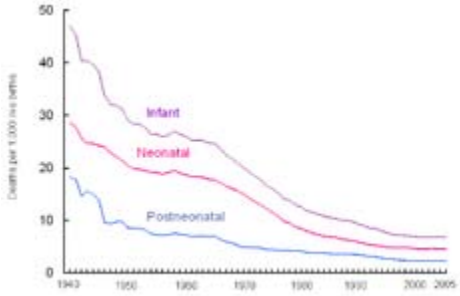
Scatter plot of the natural logarithm of the crude death rate against the natural log of per capita real GDP. The slope of the trend line is the elasticity of the crude death rate with respect to per capita real income. It indicates that a 10% increase in per capita real income is associated with a 1.5% decrease in the crude death rate.

According to Jean Ziegler (the United Nations Special Rapporteur on the Right to Food for 2000 to March 2008), mortality due to malnutrition accounted for 58% of the total mortality in 2006: "In the world, approximately 62 millions people, all causes of death combined, die each year. In 2006, more than 36 millions died of hunger or diseases due to deficiencies in micronutrients".

Of the roughly 150,000 people who die each day across the globe, about two thirds -- 100,000 per day -- die of age-related causes. In industrialized nations, the proportion is much higher, reaching 90%.

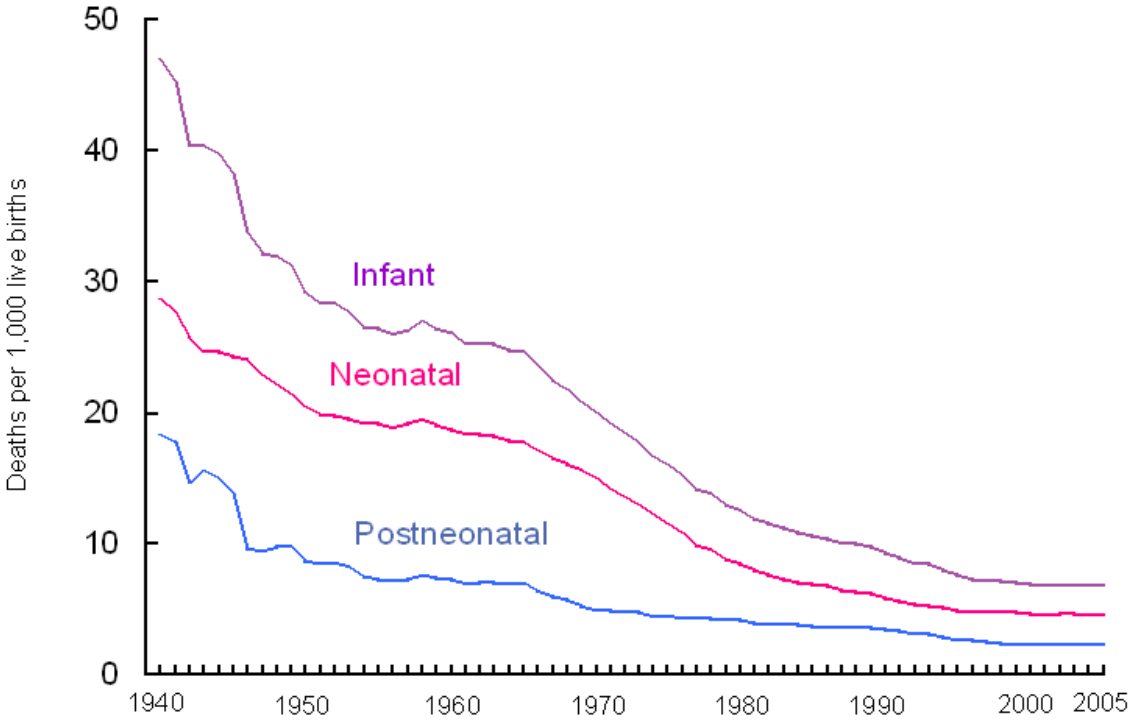
Perinatal mortality

Perinatal mortality



Infant, neonatal, and postneonatal mortality rates: United States, 1940-2005

DiseasesDB 24405



Perinatal mortality (PNM), also **perinatal death**, refers to the death of a fetus or neonate and is the basis to calculate the **perinatal mortality rate**. Variations in the precise definition of the perinatal mortality exist specifically concerning the issue of inclusion or exclusion of early fetal and late neonatal fatalities. The World Health Organization defines perinatal mortality as the "number of stillbirths and deaths in the first week of life per 1,000 live births", but other definitions have been used.

Causes

Preterm birth is the most common cause of perinatal mortality, causing almost 30 percent of neonatal deaths. Infant respiratory distress syndrome, in turn, is the leading cause of death in preterm infants, affecting about 1% of newborn infants. Birth defects cause about 21 percent of neonatal death.

Fetal mortality

Fetal mortality refers to stillbirths or fetal death. It encompasses any death of a fetus after 20 weeks of gestation or 500 gm. In some definitions of the PNM **early fetal mortality** (week 20-27 gestation) is not included, and the PNM may only include **late fetal death** and neonatal death. Fetal death can also be divided into death prior to labor, **antenatal (antepartum) death**, and death during labor, **intranatal (intrapartum) death**.

Neonatal mortality

Early neonatal mortality refers to a death of a live-born baby within the first seven days of life, while **late neonatal mortality** covers the time after 7 days until before 28 days. The sum of these two represents the neonatal mortality. Some definitions of the PNM include only the early neonatal mortality. Neonatal mortality is affected by the quality of in-hospital care for the neonate. Neonatal mortality and postneonatal mortality (covering the remaining 11 months of the first year of life) are reflected in the **Infant Mortality Rate**.

Perinatal Mortality Rate

The PNMR refers to the number of perinatal deaths per 1,000 total births. It is usually reported on an annual basis. It is a major marker to assess the quality of health care delivery. Comparisons between different rates may be hampered by varying definitions, registration bias, and differences in the underlying risks of the populations.

PNMRs vary widely and may be below 10 for certain developed countries and more than 10 times higher in developing countries. The WHO has not published contemporary data.

Maternal death

Maternal death	
ICD-10	O95
ICD-9	646.9

Maternal death, or **maternal mortality**, also "obstetrical death" is the death of a woman during or shortly after a pregnancy. In 2010, researchers from the University of Washington and the University of Queensland in Brisbane, Australia, estimated global maternal mortality in 2008 at 342,900 (down from 526,300 in 1980), of which less than 1% occurred in the developed world. However, most of these deaths have been medically preventable for decades, as treatments to avoid such deaths have been well-known since the 1950s.

Maternal Mortality definition

According to the World Health Organization, "A maternal death is defined as the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes."

Generally there is a distinction between a **direct maternal death** that is the result of a complication of the pregnancy, delivery, or their management, and an **indirect maternal death** that is a pregnancy-related death in a patient with a preexisting or newly developed health problem. Other fatalities during but unrelated to a pregnancy are termed *accidental*, *incidental*, or **nonobstetrical** maternal deaths.

Maternal mortality is a sentinel event to assess the quality of a health care system. However, a number of issues need to be recognized. First of all, the WHO definition is one of many; other definitions may also include accidental and incidental causes. Cases with "incidental causes" include deaths secondary to violence against women that may be related to the pregnancy and be affected by the socioeconomic and cultural environment. Also, it has been reported that about 10% of maternal deaths may occur late, that is after 42 days after a termination or delivery, thus, some definitions extend the time period of observation to one year after the end of the gestation. Further, it is well recognized that maternal mortality numbers are often significantly underreported.

Reducing the maternal mortality by three quarters between 1990 and 2015 is a specific part of *Goal 5 -Improving Maternal Health* - of the eight Millennium Development Goals; its progress is monitored at mdgmonitor.org

Major causes

The major causes of maternal death are bacterial infection, variants of gestational hypertension including pre-eclampsia and HELLP syndrome, obstetrical hemorrhage, ectopic pregnancy, puerperal sepsis (childbed fever), amniotic fluid embolism, uterine rupture and complications of unsafe or unsanitary abortions. Lesser known causes of maternal death include renal failure, cardiac failure, and hyperemesis gravidarum.

As stated by the 2005 World Health Organization report "Make Every Mother and Child Count" they are: severe bleeding/hemorrhage (25%), infections (13%), unsafe abortions (13%), eclampsia (12%), obstructed labour (8%), other direct causes (8%), and indirect causes (20%). Indirect causes such as malaria, anaemia, HIV/AIDS and cardiovascular disease, complicate pregnancy or are aggravated by it.

Forty-five percent of postpartum deaths occur within 24 hours. Over 90% of maternal deaths occur in developing countries. In comparison, pregnancy-associated homicide accounts for 2 to 10 deaths per 100000 live births, possibly substantially higher due to underreporting.

In developed countries, the most common cause of maternal death is obstetrical hemorrhage, followed by deep vein thrombosis.

Maternal Mortality Ratio (MMR)

Maternal Mortality Ratio is the ratio of the number of maternal deaths per 100,000 live births. The MMR is used as a measure of the quality of a health care system. Sierra Leone has the highest maternal death rate at 2,000, and Afghanistan has the second highest maternal death rate at 1900 maternal deaths per 100,000 live births, reported by the UN based on 2000 figures. According to the Central Asia Health Review, Afghanistan's maternal mortality rate was 1,600 in 2007. Lowest rates included Ireland at 0 per 100,000 and Austria at 4 per 100,000. In the United States, the maternal death rate was 11 maternal deaths per 100,000 live births in 2005. This rose to 13.3 per 100,000 in 2006. "Lifetime risk of maternal death" accounts for number of pregnancies and risk. In sub-Saharan Africa the lifetime risk of maternal death is 1 in 16, for developed nations only 1 in 2,800.

In 2003, the WHO, UNICEF and UNFPA produced a report with statistics gathered from 2000. The world average per 100,000 was 400, the average for developed regions was 20, and for developing regions 440. The worst countries were: Sierra Leone (2,000), Afghanistan (1,900), Malawi (1,800), Angola (1,700), Niger (1,600), Tanzania (1,500), Rwanda (1,400), Mali (1,200), Somalia, Zimbabwe, Chad, Central African Republic, Guinea Bissau (1,100 each), Mozambique, Burkina Faso, Burundi, and Mauritania (1,000 each).

Associated risk factors

High rates of maternal deaths occur in the same countries that have high rates of infant mortality reflecting generally poor nutrition and medical care.

Low birth weight of the child is correlated with maternal death from cardiovascular disease. Subtracting one pound of infant birth weight is correlated with the doubling of the risk of maternal death. Conversely, heavier child birth weight is correlated with lower risk of maternal death.

Another issue that is associated with maternal mortality is the distance of traveling to the nearest clinic to receive proper care. In developing nations, as well as rural areas, this is especially true. Traveling to and back from the clinic is very difficult and costly, especially to poor families when time could have been used for working and providing incomes. Even so, the nearest clinic may not provide decent care because of the lack of proper staff and equipment such as ones in the Guatemalan highlands.

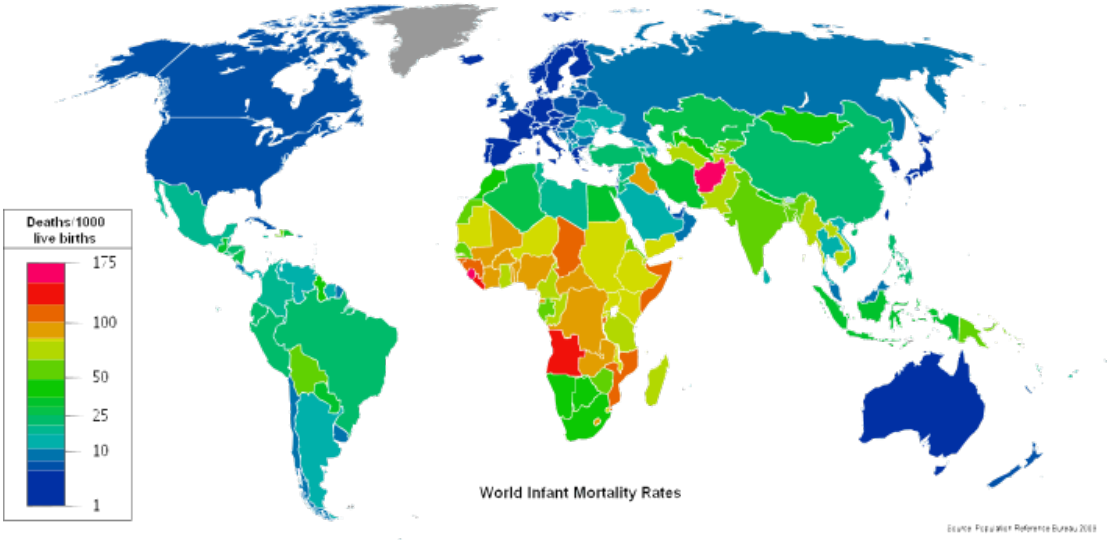
Maternal death rates in the 20th century

The death rate for women giving birth plummeted in the 20th century.

The historical level of maternal deaths is probably around 1 in 100 births. Mortality rates reached very high levels in maternity institutions in the 1800s, sometimes climbing to 40 percent of birthgiving women. At the beginning of the 1900s, maternal death rates were around 1 in 100 for live births. The number in 2005 in the United States is 11 in 100,000, a decline by two orders of magnitude, although that figure has begun to rise in recent years, having nearly tripled over the past decade in California. For the United States, 11 in 100,000 is one of the lowest estimates. Maternal deaths in the United States range up to 17 per 100,000 live births.

The decline in maternal deaths has been due largely to improved asepsis, fluid management and blood transfusion, and better prenatal care. Recommendations for reducing maternal mortality include access to health care, access to family planning services, and emergency obstetric care, funding and intrapartum care.

Infant mortality



World infant mortality rates in 2008

Infant mortality is defined as the number of infant deaths (one year of age or younger) per 1000 live births. Traditionally, the most common cause worldwide was dehydration from diarrhea. However, the spreading information about Oral Re-hydration Solution (a mixture of salts, sugar, and water) to mothers around the world has decreased the rate of children dying from dehydration. Currently, the most common cause is pneumonia. Other causes of infant mortality include: malnutrition, malaria, congenital malformation, infection and SIDS.

Infanticide, child abuse, child abandonment, and neglect also contribute to a lesser extent. Related statistical categories:

- *Perinatal mortality* only includes deaths between the foetal viability (22 weeks gestation) and the end of the 7th day after delivery.
- *Neonatal mortality* only includes deaths in the first 28 days of life.
- *Postneonatal mortality* only includes deaths after 28 days of life but before one year.
- *Child mortality* includes deaths within the first five years after birth.

World historical and predicted infant mortality rates per 1,000 births (1950–2050)
UN, medium variant, 2008 rev.

Years	Rate	Years	Rate
1950–1955	152	2000–2005	52
1955–1960	136	2005–2010	47
1960–1965	116	2010–2015	43
1965–1970	100	2015–2020	40
1970–1975	91	2020–2025	37
1975–1980	83	2025–2030	34
1980–1985	74	2030–2035	31
1985–1990	65	2035–2040	28
1990–1995	61	2040–2045	25
1995–2000	57	2045–2050	23

Cemetery at Cades Cove with three graves of infants born to the same parents in 1916, 1917 and 1918

Infant mortality rate

Infant mortality rate (IMR) indicates the number of deaths of babies under one year of age per 1,000 live births. The rate in a given region, therefore, is the total number of newborns dying under one year of age divided by the total number of live births during the year, then all multiplied by 1,000. The infant mortality rate is also called the infant death rate (per 1,000 live births).

Historically, infant mortality claimed a considerable percentage of children born, in the 1850s in america it was estamated to be as 216.8 per 1,000 for whites and 340.0 for african americans but rates have significantly declined in the West in modern times. This has been mainly due to improvements in basic health care, though high-technology medical advances have also helped. Infant mortality rate is commonly included as a part of standard of living evaluations in economics.

Comparing infant mortality rates

The infant mortality rate correlates very weakly with, and is among the best predictors of, state failure. IMR is therefore also a useful indicator of a country's level of health or development, and is a component of the physical quality of life index. However, the method of calculating IMR often varies widely between countries, and is based on how

they define a live birth and how many premature infants are born in the country. The World Health Organization (WHO) defines a live birth as any born human being who demonstrates independent signs of life, including breathing, voluntary muscle movement, or heartbeat. Many countries, however, including certain European states and Japan, only count as live births cases where an infant breathes at birth, which makes their reported IMR numbers somewhat lower and raises their rates of perinatal mortality.

The exclusion of any risk infants from the denominator or numerator in reported IMRs can be problematic for comparisons. Many countries, including the United States, Sweden or Germany, count an infant exhibiting any sign of life as alive, no matter the month of gestation or the size, but according to United States Centers for Disease Control (CDC) researchers, some other countries differ in these practices. All of the countries named adopted the WHO definitions in the late 1980s or early 1990s, which are used throughout the European Union. However, in 2009, the US CDC issued a report that stated that the American rates of infant mortality were affected by the United States' high rates of premature babies compared to European countries. It also outlined the differences in reporting requirements between the United States and Europe, noting that France, the Czech Republic, Ireland, the Netherlands, and Poland do not report all live births of babies under 500 g and/or 22 weeks of gestation. The report concluded, however, that the differences in reporting are unlikely to be the primary explanation for the United States' relatively low international ranking.

Another well-documented example also illustrates this problem. Until the 1990s, Russia and the Soviet Union did not count, as a live birth or as an infant death, extremely premature infants (less than 1,000 g, less than 28 weeks gestational age, or less than 35 cm in length) that were born alive (breathed, had a heartbeat, or exhibited voluntary muscle movement) but failed to survive for at least seven days. Although such extremely premature infants typically accounted for only about 0.005% of all live-born children, their exclusion from both the numerator and the denominator in the reported IMR led to an estimated 22%-25% lower reported IMR. In some cases, too, perhaps because hospitals or regional health departments were held accountable for lowering the IMR in their catchment area, infant deaths that occurred in the 12th month were "transferred" statistically to the 13th month (i.e., the second year of life), and thus no longer classified as an infant death.

UNICEF uses a statistical methodology to account for reporting differences among countries:

> “ UNICEF compiles infant mortality country estimates derived from all sources and methods of estimation obtained either from standard reports, direct estimation from micro data sets, or from UNICEF's yearly exercise. In order to sort out differences between estimates produced from different sources, with different methods, UNICEF developed, in coordination with WHO, the WB and UNSD, an estimation methodology that minimizes the errors embodied in each estimate and harmonize trends along time. Since the estimates are not ”

necessarily the exact values used as input for the model, they are often not recognized as the official IMR estimates used at the country level. However, as mentioned before, these estimates minimize errors and maximize the consistency of trends along time.

Another challenge to comparability is the practice of counting frail or premature infants who die before the normal due date as miscarriages (spontaneous abortions) or those who die during or immediately after childbirth as stillborn. Therefore, the quality of a country's documentation of perinatal mortality can matter greatly to the accuracy of its infant mortality statistics. This point is reinforced by the demographer Ansley Coale, who finds dubiously high ratios of reported stillbirths to infant deaths in Hong Kong and Japan in the first 24 hours after birth, a pattern that is consistent with the high recorded sex ratios at birth in those countries. It suggests not only that many female infants who die in the first 24 hours are misreported as stillbirths rather than infant deaths, but also that those countries do not follow WHO recommendations for the reporting of live births and infant deaths.

Another seemingly paradoxical finding is that when countries with poor medical services introduce new medical centers and services, instead of declining the reported IMRs often increase for a time. This is mainly because improvement in access to medical care is often accompanied by improvement in the registration of births and deaths. Deaths that might have occurred in a remote or rural area, and not been reported to the government, might now be reported by the new medical personnel or facilities. Thus, even if the new health services reduce the actual IMR, the reported IMR may increase.

Infant mortality rate in war

In most cases, war-affected areas will experience a significant increase in infant mortality rates. The primary causes of the increase are external factors such as murder and abuse. However, many other significant factors influence infant mortality rates in war-torn areas. Health care systems in developing countries in the midst of war often collapse. Attaining basic medical supplies and care becomes increasingly difficult. During the Yugoslav Wars in the 1990s Bosnia experienced a 60% decrease in child immunizations. Preventable diseases can quickly become epidemic given the medical conditions during war.

Many developing countries rely on foreign aid for basic nutrition. Transport of aid becomes significantly more difficult in times of war. In most situations the average weight of a population will drop substantially. Expecting mothers are affected even more by lack of access to food and water. During the Yugoslav Wars in Bosnia the number of premature babies born increased and the average birth weight decreased.

There have been several instances in recent years of systematic rape as a weapon of war. Women who become pregnant as a result of war rape face even more significant challenges in bearing a healthy child. Studies suggest that women who experience sexual

violence before or during pregnancy are more likely to experience infant death in their children. Causes of infant mortality in abused women range from physical side effects of the initial trauma to psychological effects that lead to poor adjustment to society. Many women who became pregnant by rape in Bosnia were isolated from their hometowns making life after childbirth exponentially more difficult.

Global infant mortality trends

For the world, and for both Less Developed Countries (LDCs) and More Developed Countries (MDCs), IMR declined significantly between 1960 and 2001. According to the Save the Children State of the World's Mothers report, the world infant mortality rate declined from 126 in 1960 to 57 in 2001.

However, IMR was, and remains, higher in LDCs. In 2001, the Infant Mortality Rate for Less Developed Countries (91) was about 10 times as large as it was for More Developed Countries (8). For Least Developed Countries, the Infant Mortality Rate is 17 times as high as it is for More Developed Countries. Also, while both LDCs and MDCs made dramatic reductions in infant mortality rates, reductions among less developed countries are, on average, much less than those among the more developed countries.

Infant mortality rate in countries

Nikolai Yaroshenko. *Funeral of Firstborn*, 1893

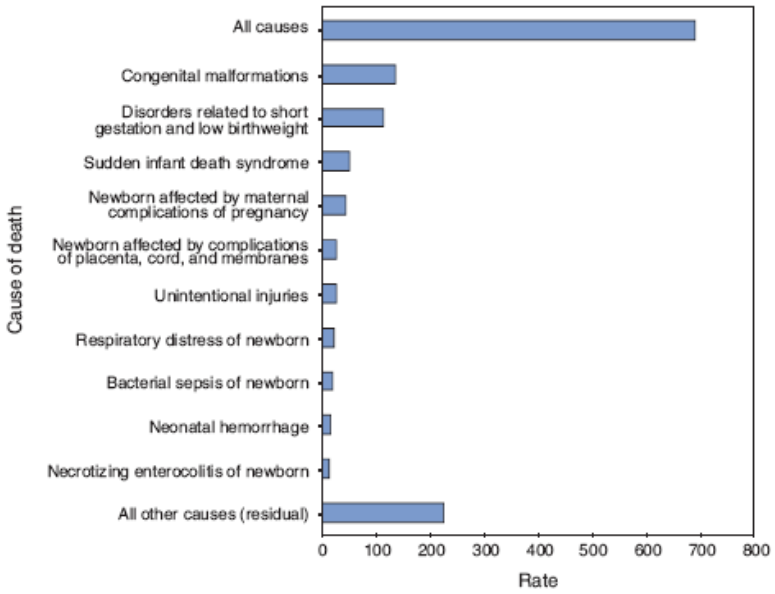
Nearly two orders of magnitude separate countries with the highest and lowest reported infant mortality rates. The top and bottom five countries by this measure (taken from The World Factbook's 2009 estimates) are shown below.

Rank	Country	Infant mortality rate (deaths/1,000 live births)
1	Angola	180.21
2	Sierra Leone	154.43
3	Afghanistan	151.95
4	Liberia	138.24
5	Niger	116.66
219	Hong Kong	2.92
220	Japan	2.79
221	Sweden	2.75
222	Bermuda	2.46
223	Singapore	2.31

Afghanistan's infant mortality rate is expected to improved by at least 60% in the next ten years due to billions of dollars of international aid.

United States

In the United States, infant mortality is 630 per 100,000 live births or 6.3 per 1000 live births.



Infant mortality rates in the United States per 100,000 live births for 10 leading causes of, 2005

The infant mortality rate for European Americans was 5.7 per 1000 births in 2003-05. For African Americans it was 13.6 per 1000, and for Hispanic Americans it was 5.6 per 1000. Overall, the infant mortality rate for the United States was 6.9 per 1000 in 2003-05.

Infant Mortality Rate by State (2005)

State	Infant mortality Rate per 1000 births
Alabama	8.96
Alaska	6.45
Arizona	6.69
Arkansas	8.29
California	5.22
Colorado	6.27
Connecticut	5.53
Delaware	9.03
District of Columbia	14.1
Florida	7.24
Georgia	8.35
Hawaii	6.67
Idaho	6.12
Illinois	7.53
Indiana	7.87
Iowa	5.40
Kansas	7.12
Kentucky	6.79
Louisiana	9.79
Maine	5.87
Maryland	8.00
Massachusetts	4.89
Michigan	8.02
Minnesota	4.78
Mississippi	10.74
Missouri	7.63
Montana	6.35
Nebraska	5.89
Nevada	5.86
New Hampshire	5.05
New Jersey	5.44
New Mexico	6.13

New York	6.02
North Carolina	8.85
North Dakota	6.35
Ohio	7.82
Oklahoma	7.86
Oregon	5.68
Pennsylvania	7.30
Rhode Island	6.20
South Carolina	9.03
South Dakota	7.18
Tennessee	8.87
Texas	6.45
Utah	4.92
Vermont	5.37
Virginia	7.50
Washington	5.39
West Virginia	8.1
Wisconsin	6.34
Wyoming	6.95

Child mortality

Child mortality, also known as **under-5 mortality**, refers to the death of infants and children under the age of five. In 2009, 8.1 million children under five died, down from 8.8 million in 2008, and 12.4 million in 1990. About half of child deaths occur in Africa. Approximately 60 countries make up 94% of under five child deaths. Reduction of child mortality is the fourth of the United Nations' Millennium Development Goals.

Causes

According to UNICEF, most child deaths (and 70% in developing countries) result from one the following five causes or a combination thereof:

- acute respiratory infections
- diarrhea
- measles
- malaria
- malnutrition

Two-thirds of deaths are preventable. Malnutrition and the lack of safe water and sanitation contribute to half of all these children's deaths. Research and experience show that most of the children who die each year could be saved by low-tech, evidence-based, cost-effective measures such as vaccines, antibiotics, micronutrient supplementation, insecticide-treated bed nets, improved family care and breastfeeding practices , and oral rehydration therapy. In addition to providing vaccines and antibiotics to children, education could also be provided to mothers about how they can make simple changes to living conditions such as improving hygiene in order to increase the health of their children. Mothers who are educated will also have increased confidence in the ability to take care of their children, therefore providing a healthier relationship and environment for them.

Rate

The **child mortality rate** or **under-5 mortality rate** is the number of children who die by the age of five, per thousand live births. In 2009, the world average was 60 (6.0%), down from 68 (6.8%) in 2007 and 89 (8.9%) in 1990. In 2006, the average in developing countries was 79 (down from 103 in 1990), whereas the average in industrialized countries was 6 (down from 10 in 1990). One in eight children in Sub-Saharan Africa die before their fifth birthday. The biggest improvement between 1990 and 2006 was in Latin America and the Caribbean, which cut their child mortality rates by 50%. The world's child mortality rate has dropped by over 60% since 1960.

A child in Sierra Leone, which has the world's highest child mortality rate (262 in 2007) is about 87 times more likely to die than one born in Sweden (with a rate of 3). According to a Save the Children paper, there are huge disparities in the under-five mortality rate between rich and poor households in developing countries. For example, children from the poorest households in India are three times more likely to die before their fifth birthday than those from the richest households.

According to the World Health Organization, the main causes of death are pneumonia, diarrhea, malaria, measles, and HIV. Malnutrition is estimated to contribute to more than one third of all child deaths. 1 child dies every 5 seconds as a result of hunger - 700 every hour - 16 000 each day - 6 million each year - 60% of all child deaths (2002-2008 estimates).

Highest rates in the world

In 2007, there were 37 countries in which at least 10% of children under five died, down from 41 in 2006. All were in Africa, except for Afghanistan. Seven of the 37 had higher rates of child mortality than in 1990. The highest 20 were:

In deaths per thousand

1. Sierra Leone - 262
2. Afghanistan - 257

3. Chad - 209
4. Equatorial Guinea - 206
5. Guinea-Bissau - 198
6. Mali - 196
7. Burkina Faso - 191
8. Nigeria - 189
9. Rwanda - 181
10. Burundi - 180
11. Niger - 176
12. Central African Republic - 172
13. Zambia - 170
14. Mozambique - 168
15. Democratic Republic of the Congo - 161
16. Angola - 158
17. Guinea - 150
18. Cameroon - 148
19. Somalia - 142
20. Liberia - 133

Standardised mortality rate

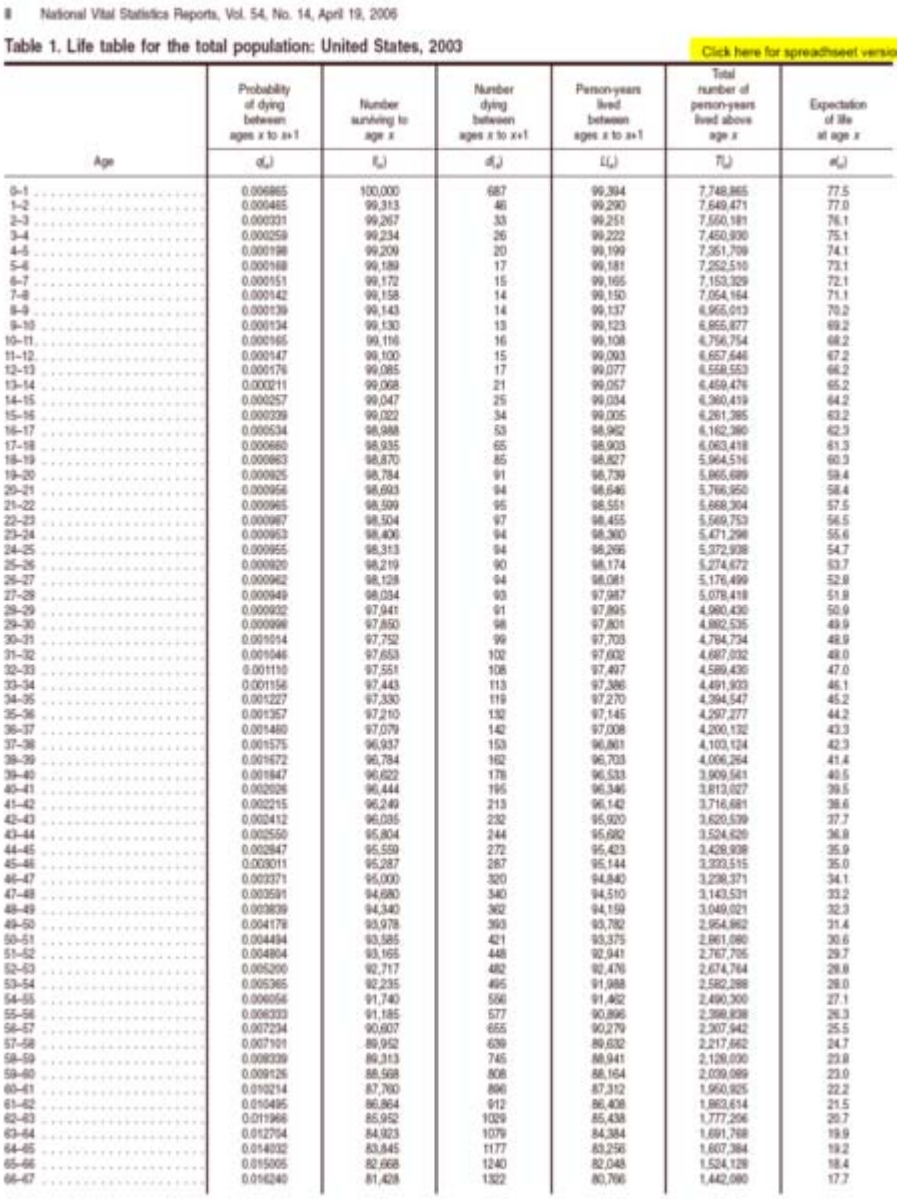
Standardized mortality ratio (indirect age adjustment) tells how many persons, per thousand of the population, will die in a given year and what the causes of death will be. Such statistics have many uses.

- Life insurance companies periodically update their premiums based on the mortality rate, adjusted for age.
- Medical researchers can track disease-related deaths and shift focus and funding to address increasing or decreasing risks.
- Organizations, both non- and for-profit, can utilize such statistics to justify their missions.
- Regarding occupational uses:

Mortality tables are also often used when numbers of deaths for each age-specific stratum are not available. It is also used to study mortaliy rate in an occupationally exposed population: Do people who work in a certain industry, such as mining or construction, have a higher mortality than people of the same age in the general population? Is an additional risk associated with that occupation? To answer the question of whether a population of miners has a higher mortality than we would expect in a similar population that is not engaged in mining, the age-specific rates for such a known population, such as all men of the same age, are applied to each age group in the population of interest. This will yield the number of deaths expected in each age group in the population of interest, if this population had had the mortality experience of the known population. Thus, for

each age group, the number of deaths expected is calculated, and these numbers are totaled. The numbers of deaths that were actually observed in that population are also calculated and totaled. The ratio of the total number of deaths actually observed to the total number of deaths expected, if the population of interest had had the mortality experience of the known population, is then calculated. This ratio is called the standardized mortality ratio (SMR). The SMR is defined as follows: SMR = (Observed no. of deaths per year)/(Expected no. of deaths per year).

Life table



8 National Vital Statistics Reports, Vol. 54, No. 14, April 19, 2006

Table 1. Life table for the total population: United States, 2003

Click here for spreadhseet versio

Age	Probability of dying between ages x to $x+1$ q_x	Number surviving to age x l_x	Number dying between ages x to $x+1$ d_x	Person-years lived between ages x to $x+1$ L_x	Total number of person-years lived above age x T_x	Expectation of life at age x e_x
0–1	0.006865	100,000	687	99,394	7,748,865	77.5
1–2	0.000485	99,313	48	99,290	7,649,471	77.0
2–3	0.000331	99,267	33	99,251	7,550,181	76.1
3–4	0.000259	99,234	26	99,222	7,450,930	75.1
4–5	0.000198	99,209	20	99,199	7,351,709	74.1
5–6	0.000168	99,189	17	99,181	7,252,510	73.1
6–7	0.000151	99,172	15	99,165	7,153,329	72.1
7–8	0.000142	99,158	14	99,150	7,054,164	71.1
8–9	0.000139	99,143	14	99,137	6,955,013	70.2
9–10	0.000134	99,130	13	99,123	6,855,877	69.2
10–11	0.000165	99,116	16	99,108	6,756,754	68.2
11–12	0.000147	99,100	15	99,093	6,657,646	67.2
12–13	0.000176	99,085	17	99,077	6,558,553	66.2
13–14	0.000211	99,068	21	99,057	6,459,476	65.2
14–15	0.000257	99,047	25	99,034	6,360,419	64.2
15–16	0.000339	99,022	34	99,005	6,261,385	63.2
16–17	0.000534	98,988	53	98,962	6,162,380	62.3
17–18	0.000660	98,935	65	98,903	6,063,418	61.3
18–19	0.000863	98,870	85	98,827	5,964,516	60.3
19–20	0.000925	98,784	91	98,739	5,865,689	59.4
20–21	0.000956	98,693	94	98,646	5,766,950	58.4
21–22	0.000965	98,599	95	98,551	5,668,304	57.5
22–23	0.000987	98,504	97	98,455	5,569,753	56.5
23–24	0.000953	98,406	94	98,360	5,471,298	55.6
24–25	0.000955	98,313	94	98,266	5,372,938	54.7
25–26	0.000920	98,219	90	98,174	5,274,672	53.7
26–27	0.000962	98,128	94	98,081	5,176,499	52.8
27–28	0.000949	98,034	93	97,987	5,078,418	51.8
28–29	0.000932	97,941	91	97,895	4,980,430	50.9
29–30	0.000998	97,850	98	97,801	4,882,535	49.9
30–31	0.001014	97,752	99	97,703	4,784,734	48.9
31–32	0.001046	97,653	102	97,602	4,687,032	48.0
32–33	0.001110	97,551	108	97,497	4,589,430	47.0
33–34	0.001156	97,443	113	97,386	4,491,933	46.1
34–35	0.001227	97,330	119	97,270	4,394,547	45.2
35–36	0.001357	97,210	132	97,145	4,297,277	44.2
36–37	0.001460	97,079	142	97,008	4,200,132	43.3
37–38	0.001575	96,937	153	96,861	4,103,124	42.3
38–39	0.001672	96,784	162	96,703	4,006,264	41.4
39–40	0.001847	96,622	178	96,533	3,909,561	40.5
40–41	0.002026	96,444	195	96,346	3,813,027	39.5
41–42	0.002215	96,249	213	96,142	3,716,681	38.6
42–43	0.002412	96,035	232	95,920	3,620,539	37.7
43–44	0.002550	95,804	244	95,682	3,524,620	36.8
44–45	0.002847	95,559	272	95,423	3,428,938	35.9
45–46	0.003011	95,287	287	95,144	3,333,515	35.0
46–47	0.003371	95,000	320	94,840	3,238,371	34.1
47–48	0.003591	94,680	340	94,510	3,143,531	33.2
48–49	0.003839	94,340	362	94,159	3,049,021	32.3
49–50	0.004179	93,978	393	93,782	2,954,862	31.4
50–51	0.004494	93,585	421	93,375	2,861,080	30.6
51–52	0.004804	93,165	448	92,941	2,767,705	29.7
52–53	0.005200	92,717	482	92,476	2,674,764	28.8
53–54	0.005365	92,235	495	91,988	2,582,288	28.0
54–55	0.006056	91,740	556	91,462	2,490,300	27.1
55–56	0.006333	91,185	577	90,896	2,398,838	26.3
56–57	0.007234	90,607	655	90,279	2,307,942	25.5
57–58	0.007101	89,952	639	89,632	2,217,662	24.7
58–59	0.008339	89,313	745	88,941	2,128,030	23.8
59–60	0.009126	88,568	808	88,164	2,039,089	23.0
60–61	0.010214	87,760	896	87,312	1,950,925	22.2
61–62	0.010495	86,864	912	86,408	1,863,614	21.5
62–63	0.011966	85,952	1029	85,438	1,777,206	20.7
63–64	0.012704	84,923	1079	84,384	1,691,768	19.9
64–65	0.014032	83,845	1177	83,256	1,607,384	19.2
65–66	0.015005	82,668	1240	82,048	1,524,128	18.4
66–67	0.016240	81,428	1322	80,766	1,442,080	17.7

2003 US mortality table, Table 1, Page 1

In actuarial science, a **life table** (also called a **mortality table** or **actuarial table**) is a table which shows, for each age, what the probability is that a person of that age will die before his next birthday. From this starting point, a number of statistics can be derived and thus also included in the table:

- the probability of surviving any particular year of age
- remaining life expectancy for people at different ages
- the proportion of the original birth cohort still alive
- estimates of a cohort's longevity characteristics.

Life tables are usually constructed separately for men and for women because of their substantially different mortality rates. Other characteristics can also be used to distinguish different risks, such as smoking status, occupation, and socio-economic class.

Life tables can be extended to include other information in addition to mortality, for instance health information to calculate health expectancy. Health expectancies, of which disability-free life expectancy (DFLE) and Healthy Life Years (HLY) are the best-known examples, are the remaining number of years a person can expect to live in a specific health state, such as free of disability. Two types of life tables are used to divide the life expectancy into life spent in various states: 1) multi-state life tables (also known as increment-decrement life tables) based on transition rates in and out of the different states and to death, and 2) prevalence-based life tables (also known as the Sullivan method) based on external information on the proportion in each state. Life tables can also be extended to show life expectancies in different labor force states or marital status states.

Life tables are also used extensively in biology and epidemiology. The concept is also of importance in product life cycle management.

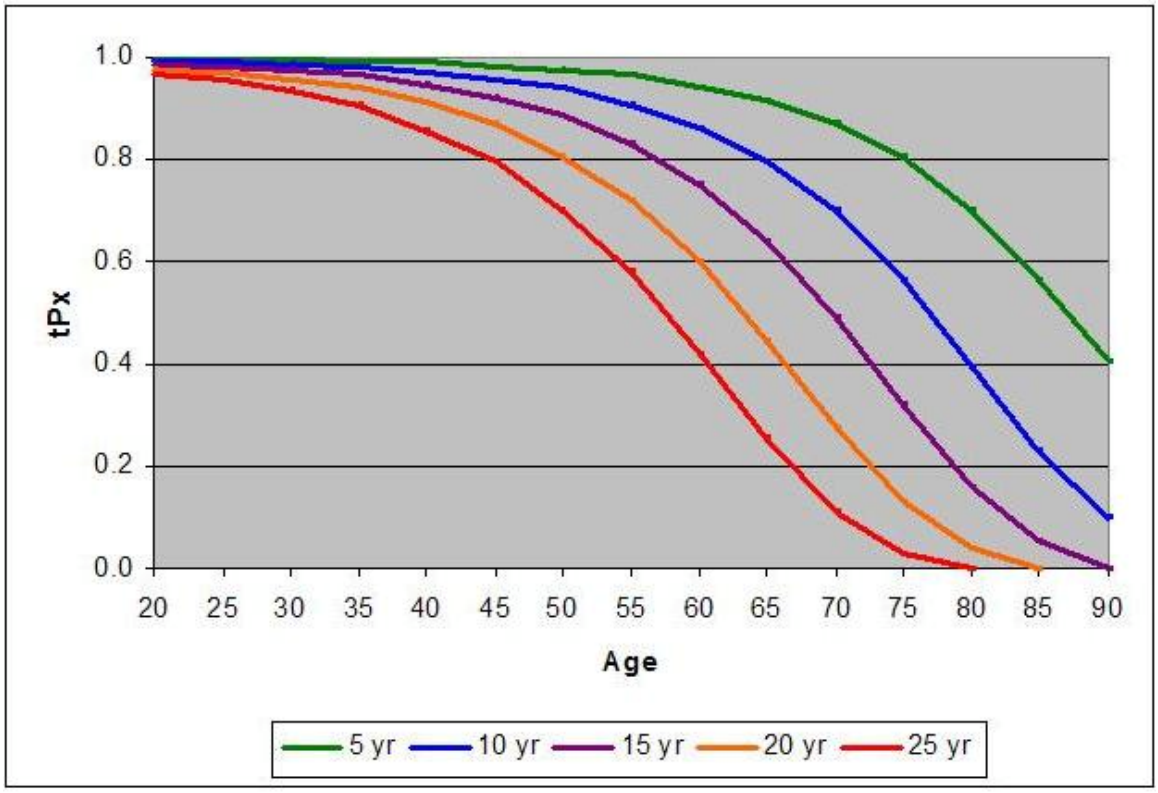
Insurance applications

In order to price insurance products, and ensure the solvency of insurance companies through adequate reserves, actuaries must develop projections of future insured events (such as death, sickness, and disability). To do this, actuaries develop mathematical models of the rates and timing of the events. They do this by studying the incidence of these events in the recent past, and sometimes developing expectations of how these past events will change over time (for example, whether the progressive reductions in mortality rates in the past will continue) and deriving expected rates of such events in the future, usually based on the age or other relevant characteristics of the population. These are called mortality tables if they show death rates, and morbidity tables if they show various types of sickness or disability rates.

The availability of computers and the proliferation of data gathering about individuals has made possible calculations that are more voluminous and intensive than those used in the past (i.e. they crunch more numbers) and it is more common to attempt to provide different tables for different uses, and to factor in a range of non-traditional behaviors (e.g. gambling, debt load) into specialized calculations utilized by some institutions for

evaluating risk. This is particularly the case in non-life insurance (eg the pricing of motor insurance can allow for a large number of risk factors, which requires a correspondingly complex table of expected claim rates).

The mathematics



${}_tp_x$ chart from Table 1. Life table for the total population: United States, 2003, Page 8

The basic algebra used in life tables is as follows.

- q_x**:** the probability that someone aged exactly $x$will die before reaching age $(x+1)$.
- p_x**:** the probability that someone aged exactly $x$will survive to age $(x+1)$.

$$p_x = 1 - q_x$$

- l_x**:** the number of people who survive to age x

  note that this is based on a starting point of $l_0$lives, typically taken as 100,000

$$l_{x+1} = l_x \cdot (1 - q_x) = l_x \cdot p_x$$

$$\frac{l_{x+1}}{l_x} = p_x$$

- d_x**:** the number of people who die aged $x$last birthday

$$d_x = l_x - l_{x+1} = l_x \cdot (1 - p_x) = l_x \cdot q_x$$

- ${}_tp_x$**:** the probability that someone aged exactly $x$will survive for $t$more years, i.e. live up to at least age $x + t$years

$${}_tp_x = \frac{l_{x+t}}{l_x}$$

- ${}_{t|k}q_x$**:** the probability that someone aged exactly $x$will survive for $t$more years, then die within the following $k$years

$${}_{t|k}q_x = {}_tp_x \cdot {}_kq_{x+t} = \frac{l_{x+t} - l_{x+t+k}}{l_x}$$

- μ_x : the *force of mortality*, ie the instantaneous mortality rate at age *x*, ie the number of people dying in a short interval starting at age *x*, divided by l_x and also divided by the length of the interval. Unlike q_x, the instantaneous mortality rate, μ_x, may exceed 1.

Biology

When biologists and demographers use life tables, they will normally also include fertility for each age. The extra parameter used is

- m_x**:**

Under standard international actuarial notation this symbol is referred to as the Central rate of mortality. It can be considered a form of the force of mortality, weighted by survival probability over the year of age.

Epidemiology

In epidemiology and public health, both standard life tables to calculate life expectancy and Sullivan and multistate life tables to calculate **health expectancy** are commonly used. The latter include information on health in addition to mortality.