Epidemiology is the field of medicine concerned with the study of epidemics, outbreaks of disease that affect large numbers of people. Epidemiologists, using sophisticated statistical analyses, field investigations, and complex laboratory techniques, investigate the cause of a disease, its distribution (geographic, ecological, and ethnic), method of spread, and measures for control and prevention. Epidemiological investigations once concentrated on communicable diseases such as tuberculosis, influenza, and cholera, but now also encompass cancer, heart disease, and other frequent disorders.

One of the most ambitious epidemiological research projects ever undertaken was 'The Framingham Heart Study', which was initiated in 1948 to characterise factors associated with cardiovascular disease. The Framingham Heart Study was the first of its kind. It was a 'prospective' study: it looked at a randomly chosen, representative group of people from the adult population of Framingham, Massachusetts, and collected data on them periodically over time.

The Framingham Heart Study determined many risk factors for heart disease and, at the same time, established new methods that could be effectively generalised to the study of other chronic diseases. Essentially, the science of chronic disease epidemiology, as we know it today, grew out of this work.
What is Epidemiology?

Epidemiology, literally translated from Greek, means "the study of [a] people".


Epidemiology is the study of how a disease is distributed in a population and of the factors that influence or determine this distribution.

http://www.usuhs.mil/2005/Epid_Notes_1.htm

More precisely, the Greek prefix "epi-" means "over, on, upon" and, as used in the word "epidemiology", ascribes the scientist a surveying position above people ("demos"). Indeed, in a way, epidemiologists do view populations through a microscope except that, instead of a single cell, an organ or a patient, the whole group of people is the study object.

The second quotation highlights the importance of statistics in epidemiology. Unless it is feasible to perform comprehensive surveys of populations, in which all relevant disease manifestations and outcomes can be monitored with high quality, epidemiologists are forced to rely upon samples to make conclusions about the population as a whole. The use of the word "distribution" further indicates that disease outcomes are to be regarded as random variables, and that epidemiology is the art of making sound inferences about the distribution of these random variables.
Objectives of Epidemiology

- determine the **extent** of disease in a population
- study the **natural history** and **prognosis** of a disease
- characterise the **aetiology** of a disease
- identify **risk factors** or **protective factors**
- evaluate **preventive** and **therapeutic measures**
- provide the foundation for developing **public policy and regulatory decisions**

One concern of epidemiology is the study of the frequency and distribution of health events within groups of people. For this purpose, researchers use the methodology of descriptive epidemiology which allows them to characterise health events in terms of their timing, location, and grouping.

However, epidemiologists also attempt to search for the causes of disease, or for factors that are associated with an increased risk for disease. This type of epidemiology, where we move from questions of "who," "what," "where," and "when" to answering the "how" and "why," is referred to as analytical epidemiology.

Finally, although epidemiology can be used simply as an analytical tool for studying diseases and their determinants, it also serves a more active role. Epidemiological data steer public health decision making and aid in the development and evaluation of interventions that aim to control and prevent health problems. This is the primary function of 'applied', or 'field', epidemiology.
In 1796, the English practitioner Edward Jenner carried out an experiment on one of his patients called James Phipps, an eight year old boy. Jenner made two cuts in the boy’s arm and worked into them a small amount of cowpox pus. Although the boy had the normal reaction of a slight fever, after several days he recovered. When Jenner repeated the inoculation a few weeks later, using smallpox matter, the boy remained healthy. Vaccination treatment had been born, named after the medical name for cowpox, 'vaccinia', which in turn derives from the Latin word 'vacca' for 'cow'.

In 1854, a serious cholera outbreak occurred in the Soho District of London. John Snow, a local surgeon, reasoned that if cholera was spread by a mist or miasma, as the prevailing theories suggested, then the cases should have been uniformly distributed along the streets. To see if this was the case, he plotted each cholera case on a map and also highlighted houses with multiple cases. It became apparent that the cases were not uniformly distributed, but instead formed a tight cluster around a water pump located on Broad Street. Snow went to the pump and took a water sample. Looking in his microscope, he found that the water contained bacteria he had not seen before. Snow guessed that the bacteria were responsible for the epidemic, and went back to the pump and removed the pump handle. The Broad Street cholera outbreak stopped almost overnight.
Prevalence is the proportion of individuals in a population that have a particular disease. However, even for a chronic disease, manifestations are often intermittent. Consequently, a point prevalence that is based upon a single examination at one point in time may underestimate the true frequency of a condition. If repeated or continuous assessment of individuals is possible, a better measure is the period prevalence, defined as the proportion of a population that is, or has been, affected during a specific time period. Unfortunately, the population base of a period prevalence will often be difficult to define in practice because populations change over time owing to births, deaths and migration. In such cases, either the average or the mid-interval population size may be used as the denominator of the period prevalence.

Prevalence data provide an indication of the extent of a condition in a population and may have implications for the provision of services needed in the community. On the other hand, prevalence is an appropriate measure of morbidity only for relatively stable conditions, and it may thus be unsuitable in the context of service provision for acute disorders.
In the present example, the denominator of the period prevalence for the 30 years period of interest equals the mid-interval population size, i.e. the population size after 15 years (marked by a vertical line).

\[
\hat{\pi}_d = \frac{4}{7} = 0.57 \\
\hat{\pi}_t = \frac{3}{7} = 0.43
\]
Confidence Interval

The number $X$ of diseased individuals in a sample of size $n$ follows a $\text{Bin}(n, \pi)$ distribution.

\[ KI: \hat{\pi} \pm t_{1-\alpha/2, n-1} \cdot \sqrt{\frac{\hat{\pi} \cdot (1 - \hat{\pi})}{n}} \]

yields a confidence interval for the estimate of $\pi$.

Example: 20 diseased among 35 probands

\[ \hat{\pi}_d = \frac{20}{35} = 0.57 \]

\[ 95\% \text{ KI}: 0.57 \pm 2.04 \cdot \sqrt{\frac{0.57 \cdot 0.43}{35}} = 0.57 \pm 0.17 \]
The incidence proportion of a disease is defined as the number of new cases of disease occurring in a population during a defined time period, divided by the total number of people at risk during that time period.
Incidence Proportion

\[ \hat{p} = \frac{3}{9} = 0.333 \]
Sometimes, the measurement of incidence proportions is complicated by changes in the risk population during the observation period. This difficulty is overcome by relating the number of new cases to the person-years at risk, calculated by adding together the time each individual member of the population was at risk during the time period of interest. The resulting ratio, called the 'incidence rate' of a disease, can be viewed as the "speed" at which new cases occur in a population and thus depends on the time unit used. Unfortunately, the same symbol $\gamma$ is used in the literature to denote both incidence rate and incidence proportion.

It should be noted that, once a person is classified as a case, he or she is no longer liable to become a new case and therefore should not contribute further person-years at risk. Sometimes, however, the same pathological event happens more than once to the same individual. For example, a patient may have several episodes of myocardial infarction during the time period of interest. In these circumstances, the definition of incidence is usually restricted to the first event, although sometimes (for example in the study of infectious diseases) it is more appropriate to count all episodes. When ambiguity is possible, reports should state whether "incidence" refers only to the first diagnosis or to all episodes, as this may influence its interpretation.

**Morbidity Measures**

*Incidence Rate ($\gamma$), "Risk"

\[ \gamma = \frac{A}{\sum_{i=1}^{N} T_i} \]

mathematically: (time) rate at which unaffected, randomly chosen individuals get affected
\[ \hat{\gamma} = \frac{3}{163} = 0.018 \text{ incidents per person-year} \]
Prevalence and incidence are closely related morbidity measures since the proportion of a population affected by a particular disease does, of course, depend upon the proportion of new cases that arise. However, prevalence is also a function of disease duration, and the longer it takes for patients to either recover or decease, the more acute cases will be prevalent.

New cases derive from the group of unaffected individuals which comprises a proportion $1 - \pi$ of the total population. Of these unaffected individuals, a proportion $\gamma$ enters the prevalence pool per time unit $\Delta t$. Therefore, the expected prevalence pool inflow in time $\Delta t$ equals

$$\gamma \cdot (1 - \pi) \cdot \Delta t$$

During the same time, a proportion $1/E(D)$ leaves the prevalence pool (of relative size $\pi$), assuming that the expected disease duration $E(D)$ is also measured in $\Delta t$ units. For example, if a disease would constantly last for nine months, then the prevalence pool would lose $1/9$ of its patients per month. Thus, the expected prevalence pool outflow in time $\Delta t$ equals

$$\frac{1}{E(D)} \cdot \pi \cdot \Delta t$$
Prevalence and Incidence

In a stable, closed population (i.e. without migration into, or out of, the population)

\[ \gamma \cdot (1 - \pi) \cdot \Delta t = \frac{1}{E(D)} \cdot \pi \cdot \Delta t \]

so that

\[ \frac{\pi}{1 - \pi} = \gamma \cdot E(D) \]

Through causing longer disease duration, improved medical care may increase the disease burden to society in the form of an increased prevalence.

The first equation assumes that prevalence pool inflow and outflow are at an equilibrium. Such an equilibrium will be attained after some time if the population in question is (approximately) stable in terms of both disease management and population dynamics. This means that, in addition to the incidence proportion and disease duration being constant, the population size and other, potentially disease-relevant demographic factors such as sex ratio, age structure etc. do not change.

For comparatively rare diseases, the denominator \(1 - \pi\) in the second equation is approximately equal to unity and can be neglected. This equation may therefore be translated into the following rule of thumb: "prevalence equals incidence times average disease duration". Remember that, for these considerations to be valid, incidence and disease duration have to be measured on the same time scale (e.g. years).
Typhus is caused by a bacteria-like organism, *Rickettsia prowazekii*, transmitted by the human body louse, *Pediculus humanus*. Humans return this "favor" by infecting the louse, which is also a victim of the disease, seldom surviving its attack.

A louse becomes infected with typhus by taking a blood meal from a fever-ridden, infected human. Once in the louse's gut, the rickettsiae reproduce to such enormous numbers that they cause cells in the insect's gut to rupture. The rickettsiae then are present in the faeces of the louse. Humans become infected by rubbing or scratching the lice faeces into their skin or into their mucous membranes.

Once infected, humans experience a high fever that continues for approximately two weeks. Simultaneous symptoms may include severe headaches, bronchial disturbances, and mental confusion. Indeed, typhus is from the Greek word "typhos" meaning stupor. Mortality from typhus is incredibly high under epidemic conditions, nearing 100%.

The conditions of war are perfect for typhus to explode into a raging epidemic because poverty, crowding, mass migrations, inadequate housing, and malnutrition encourage its spread. Typhus' association with war and its devastating effect continued until World War II. A potentially horrific epidemic was averted in Sicily and Italy in 1943 through a concerted delousing campaign engineered by the Allies using the then miraculous compound, DDT.
Prevalence and Incidence

Problems

- ambiguous or incorrect diagnoses, latency
- identification of highly selected cases from hospital admissions (severity, policy)
- bad recording of cases (incomplete, missing)
- variable diagnostic standards (temporal, regional)
- ambiguous definition of population base (medical, ethnic, social)
- temporal changes of disease patterns (spatial, phenotypical)
One of the important objectives of epidemiology is the identification of disease risk factors. Instead of concentrating upon descriptive measures like the disease prevalence or incidence, analytical epidemiology therefore tries to ascertain effect measures.

The most important effect measure is the relative risk, or risk ratio, that relates the incidence rate (or proportion) among people exposed to a risk factor to the incidence rate (or proportion) among non-exposed. If this ratio is smaller than unity, exposed people face a lower disease risk so that the exposure in question is in fact "protective".

One reason why the relative risk is the measure of association most often used by epidemiologists is the empirical observation that, when two risk factors for a disease act in concert, their relative risks often come close to multiplying. For example the relative lung cancer risk of smokers is about 10, that of asbestos workers is approximately 5. People exposed to both risk factors have a 50-fold higher disease risk compared to individuals exposed to neither factor.

**Effect Measures**

**Relative Risk (ρ)**

Let a population be stratified into two strata (e.g. "exposed", "not exposed") with corresponding incidence rates or proportions ("risks") $\gamma_e$ and $\gamma_n$ during the observational period.

$$\rho = \frac{\gamma_e}{\gamma_n}$$

is called the 'relative risk' under exposure.

$\rho > 1$: "risk factor", $\rho < 1$: "protective"
Relative Risk ($\rho$)

\[ \hat{\rho} = \frac{5/10}{2/10} = \frac{0.5}{0.2} = 2.50 \]
Types of Epidemiological Studies
Experimental (Interventional)
Assignment of Exposure by Investigator

Clinical Trials
- performed on single diseased individuals in a clinical setting
- evaluation of therapeutic measures (e.g. drugs)

Field Trials
- performed on single non-diseased individuals in the field
- evaluation of preventive measures (e.g. vaccination)

Community Interventions
- performed on groups of non-diseased individuals
- evaluation of preventive measures (e.g. water treatment)
Archetypal Experimental Study

[Diagram showing time series with exposed and not exposed conditions.

Time scale runs horizontally from left to right. Bars indicate data points or events within each time segment.

Exposure status is indicated on the vertical axis: exposed and not exposed.

Legend or key might be included to explain the significance of colors or symbols used in the diagram.]
Types of Epidemiological Studies
Non-Experimental (Observational)
Assignment of Exposure by Nature

Cohort Studies
- performed prospectively on non-diseased individuals of known exposure status, disease incidences are recorded

Case-Control Studies
- performed retrospectively on individuals of known disease status, exposure status is recorded

Cross-Sectional (Prevalence) Studies
- performed retrospectively on the whole population or on a representative sample, disease and exposure is recorded
Archetypal Observational Study
In 1948, the US National Heart, Lung, and Blood Institute (NHLBI) was born, and the new institute initiated a daring and far-sighted study that would dramatically change people's understanding of cardiovascular disease. The institute sent a team of researchers and medical practitioners to the town of Framingham, Massachusetts, hoping to find out why the number of Americans dying from heart disease was on the rise.

Before Framingham, epidemiologists got most of their information from death certificates and the medical records of sick people. Joseph Mountin of the US Public Health Service wanted to do a different kind of study. The incidence of heart disease had risen rapidly for two decades to become the nation's number one killer. Mountin wanted to use the methods of epidemiology to study people starting when they were healthy to see why some later have heart attacks, hardening of the arteries, and other cardiovascular problems. He felt that if researchers learned enough about how heart disease develops, it would be possible to stop the rise in the heart disease-related death rate.

Objective
To identify the common factors or characteristics that contribute to cardiovascular disease (CVD) by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.

Design
In 1948, 5209 men and women between the ages of 30 and 62 were recruited from Framingham, Massachusetts (representing 2/3 of the adult population). In 1971, another sample of 5135 men and women was established, comprising the offspring of the original cohort and their spouses.
Many scientists at the US National Institutes of Health thought that the government's money should be spent on the laboratory investigation of cardiovascular disease, not on an epidemiological study. Furthermore, many citizens of Framingham had substantial misgivings, until civic pride, the lure of free examinations, and an ironclad promise of confidentiality prevailed.

The most formidable obstacles that Framingham scientists faced were intellectual. The traditional methods of infectious-disease epidemiology - tracing contacts and poring over death certificates - did not work for a chronic, non-infectious condition like heart disease. Large numbers of patients had to be assembled and followed prospectively over an extended period. A small army of physicians and assistants needed to be persuaded to obtain, interpret, and record data in an objective way. New statistical methods had to be invented to analyse the results.

Nevertheless, all of these problems were overcome, and the study succeeded. Numerous risk factors for cardiovascular disease were identified (Framingham scientists actually introduced the term 'risk factor' into medicine), and the focus of cardiology expanded to include the prevention as well as the treatment of heart disease.

The Framingham study has continued uninterrupted through to the present day.

---

Results

Careful monitoring of the Framingham Study population has led to the identification of the major CVD risk factors:
- high blood pressure
- high blood cholesterol
- smoking
- obesity
- diabetes
- physical inactivity
and critical information upon related factors such as age, gender, and psychosocial issues.

The Framingham study has produced approximately 1200 articles in leading medical journals.
The odds of disease is the ratio of the probability of getting the disease to the probability of not getting the disease, during a given time period. For instance if 20 smokers develop lung cancer during their lifetime, and 80 do not, the odds among these 100 smokers in favour of developing lung cancer is 20/80, or 0.25, whereas the risk is 0.20.

Disease odds play an important role in epidemiology since relative risks cannot normally be estimated from retrospective case-control studies. Case-control designs, however, are a popular research tool and therefore require alternative effect measures. As we will see in the following, one such measure can be obtained using disease odds instead of disease risks.
Odds have long been the standard way of representing probabilities as used by bookmakers. This is probably the reason why the use of odds instead of risks or probabilities is much more common in Commonwealth countries than in others.

If you bet one pound and the odds is 1:x for your horse winning the race, betting would be "at fair odds" if you get x pounds plus your one pound stake back if you win, i.e. "x for one". Any higher payoff guaranteed by the bookmakers turns the bet into a good deal for you, and you should bet. If the guaranteed payoff is lower, leave it and have a drink at the bar instead!

<table>
<thead>
<tr>
<th>payoff</th>
<th>1:5</th>
<th>1:10</th>
<th>1:50</th>
<th>1:200</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-1</td>
<td>fair</td>
<td>poor</td>
<td>poor</td>
<td>poor</td>
</tr>
<tr>
<td>10-1</td>
<td>good</td>
<td>fair</td>
<td>poor</td>
<td>poor</td>
</tr>
<tr>
<td>50-1</td>
<td>good</td>
<td>good</td>
<td>fair</td>
<td>poor</td>
</tr>
<tr>
<td>200-1</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>fair</td>
</tr>
</tbody>
</table>
The odds ratio is the analogue of the relative risk, i.e. it relates the odds among exposed individuals of becoming affected in a given time period to the odds among non-exposed individuals.

Effect Measures
Odds Ratio (OR)

\[ OR = \frac{\frac{\gamma_e}{1 - \gamma_e}}{\frac{\gamma_n}{1 - \gamma_n}} \]

If risks \( \gamma_e \) and \( \gamma_n \) are "sufficiently small" for the chosen time unit, i.e. of the order a few percent, then

\[ OR = \frac{\frac{\gamma_e}{1 - \gamma_e}}{\frac{\gamma_n}{1 - \gamma_n}} \approx \frac{\gamma_e}{\gamma_n} = \rho \]
The two incidence proportions per 10 years (2/10 for exposed and 1/10 for non-exposed) are comparatively low. Therefore, the odds ratio is very similar to the relative risk.

\[
\hat{OR} = \frac{2/8}{1/9} = 2.25 \quad \hat{\rho} = \frac{2/10}{1/10} = 2.00
\]
By contrast, the incidence proportions per 30 years (5/10 for exposed and 2/10 for non-exposed) are very high so that, not surprisingly, odds ratio and relative risk for a long-term affection by the disease in question differ quite substantially.
In a prospective cohort study, the line totals a+b (number of exposed) and c+d (number of non-exposed) are fixed. Provided that these numbers are large enough, the relative numbers of incidences within the two groups provide good estimates of the respective incidence proportions, $\gamma_e$ and $\gamma_n$. Therefore, prospective cohort studies indeed allow the estimation of relative risks, $\rho$. 
In a retrospective case-control study, the column totals $a+c$ (number of cases) and $b+d$ (number of controls) are fixed. Therefore, the data cannot provide any information about incidence proportions, neither among exposed nor among non-exposed individuals. If the sample size is large enough, however, $a/c$ and $b/d$ are good estimates of the 'odds of exposure' (i.e. the ratio of the exposure and non-exposure probabilities) among affected and non-effected individuals, respectively. Since the exposure odds ratio is mathematically equivalent to the disease odds ratio, a case-control study is therefore capable of providing an estimate of the latter.

Proof:

$$\frac{A_e}{N_e - A_e} / \frac{N_n - A_n}{N_n} = \frac{A_e / (N_e - A_e)}{A_n / (N_n - A_n)}$$

$$= \frac{A_e / (N_e - A_e)}{N_e / N_n} \div \frac{A_n / (N_n - A_n)}{N_n / N_n} = \frac{\gamma_e / (1 - \gamma_e)}{\gamma_n / (1 - \gamma_n)}$$
Confidence intervals are easier to write down for the natural logarithms of effect measures, rather than for the effect measures themselves. However, owing to the strictly monotonous relationship between $x$ and $\ln(x)$, confidence intervals for the effect measures can easily obtained by taking the exponential of the upper and lower limits of their logarithms, respectively. For example, consider the following $2 \times 2$ table which resulted from a prospective cohort study:

<table>
<thead>
<tr>
<th></th>
<th>50</th>
<th>950</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>990</td>
<td></td>
</tr>
</tbody>
</table>

The 95% confidence interval for the logarithm of the relative risk $\rho$ equals

$$\ln\left(\frac{50}{1000}\right) \pm 1.96 \cdot \sqrt{\frac{1}{50} - \frac{1}{1000} + \frac{1}{10} - \frac{1}{1000}}$$

or [0.936, 2.283]. So, the 95% confidence interval for $\rho$ is demarcated by $e^{0.936}$ and $e^{2.283}$, i.e. it equals [2.550, 9.806].
If the data in the present example came from a prospective cohort study, it would be valid to estimate the relative risk of disease among exposed individuals as 2.00. Since the incidence proportion is comparatively low (only 15/300, or 5%, of individuals got affected during the study period), this estimate consequently is close to the estimated odds ratio of 2.07. Note, however, that the data do not provide sufficient evidence for a genuine association between disease and exposure since the 95% confidence intervals of both effect measure still include unity.

<table>
<thead>
<tr>
<th></th>
<th>affected</th>
<th>not affected</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>10</td>
<td>140</td>
<td>150</td>
</tr>
<tr>
<td>not exposed</td>
<td>5</td>
<td>145</td>
<td>150</td>
</tr>
<tr>
<td>total</td>
<td>15</td>
<td>285</td>
<td>300</td>
</tr>
</tbody>
</table>

\[
\hat{\rho} = \frac{a}{(a + b)} = \frac{10}{150} = 0.0667 \\
\hat{OR} = \frac{a/c}{b/d} = \frac{10/5}{140/145} = 2.07
\]

95% CI: 0.70 - 5.71

95% CI: 0.69 - 6.21
Assuming that the previous example referred to a retrospective case-control study, it would have been advisable to increase the sample size, for example, by including more cases. Under the somewhat unrealistic assumption that the exposure odds in the expanded case sample (i.e. 100:50) is the same as in the original sample (i.e. 10:5), the odds ratio estimate remains unchanged at 2.07. However, the corresponding 95% confidence interval no longer includes unity so that we may conclude that the association is "real". In other words, the increased sample size has paid off.

Since the data belong to a case-control study, no relative risk can be estimated. Indeed, ignorant use of the respective estimator would yield a result, 1.63, that deviates from the previously obtained 2.00. Since all we have done is increase the number of cases, leaving all other aspects of the disease unchanged, such a decrease in relative risk would not make sense. The reason for this discrepancy is that the estimation of a relative risk was meaningless in the first place.

\[
\hat{OR} = \frac{a}{c} \div \frac{b}{d} = \frac{100}{50} \div \frac{140}{145} = 2.07 \\
\hat{\rho} = \frac{a}{(a+b)} \div \frac{c}{(c+d)} = \frac{100}{240} \div \frac{50}{195} = 1.63
\]

95% CI: 1.37 - 3.19
Which Effect Measure?

Case-control studies do not normally allow the estimation of relative risks.

An odds ratio provides a good approximation of the relative risk for a disease if the incidence rate (over the chosen time unit) is small.
There is some inconsistency in the literature regarding the definition of attributable risks. The terminology used here mainly follows that of Kenneth J. Rothman and Sander Greenland in their 1998 text book ‘Modern Epidemiology’, published by Lippincott, Williams & Wilkins.

According to the Rothman and Greenland terminology, the aetiological fraction comprises the newly arising cases, or incidences, for which the exposure was indeed causal. This figure cannot be estimated from epidemiological data alone but requires additional pathological, physiological or biochemical data.
The 'attributable risk' is defined as the proportion of the disease risk under exposure that is due to the exposure. In many text books, this figure is called the 'attributable risk fraction', or 'rate fraction', whereas "attributable risk" would refer only to the risk difference in the numerator. In any case, the attributable risk (fraction) can be estimated from epidemiological data and serves to quantify the amount by which the disease risk of an individual is increased by the exposure.
### Attributable Risk (AR)

\[
\gamma_{e,\text{males}} = 0.50 \\
\gamma_{n,\text{males}} = 0.20 \\
\rho_{\text{males}} = \frac{0.50}{0.20} = 2.5 \\
\]

\[
\gamma_{e,\text{females}} = 0.08 \\
\gamma_{n,\text{females}} = 0.02 \\
\rho_{\text{females}} = \frac{0.08}{0.02} = 4.0 \\
\]

\[
\text{AR}_{\text{males}} = \frac{2.5 - 1.0}{2.5} = 0.60 \\
\]

\[
\text{AR}_{\text{females}} = \frac{4.0 - 1.0}{4.0} = 0.75 \\
\]

Although the disease risk of exposed males is much higher than the disease risk of exposed females, the AR is higher in females because their relative risk is higher.
The 'population attributable risk' is defined as the proportion of incidences of a disease that is attributable to the presence of the exposure in general (not in an individual case). Again, this figure is often called the 'population attributable risk fraction', or 'excess fraction', whilst 'population attributable risk' would refer to the numerator, i.e. $\gamma - \gamma_n$, alone.

The population attributable risk equals the proportion of new cases that would be avoided in a population if the exposure were eliminated. Note, however, that there is no clear relationship between this fraction, i.e. the excess fraction, and the aetiological fraction. The aetiological fraction may be high and the PAR small, for example, if the exposure occurs very early in life and invariably causes disease, but if exposed people would preferably get the disease also from other causes later on in life. On the other hand, if the exposure is not causal but is only associated with other causal factors, the aetiological fraction may be zero whilst the PAR may be close to 100% (namely if avoiding the exposure simultaneously avoids all genuinely causal factors).
Population Attributable Risk (PAR)

\[ \gamma_{e, \text{males}} = 0.50 \quad \gamma_{e, \text{females}} = 0.08 \]
\[ \gamma_{n, \text{males}} = 0.20 \quad \gamma_{n, \text{females}} = 0.02 \]
\[ \rho_{\text{males}} = 0.50 / 0.20 = 2.5 \quad \rho_{\text{females}} = 0.08 / 0.02 = 4.0 \]
\[ f_{e, \text{males}} = 0.20 \quad f_{e, \text{females}} = 0.10 \]

\[ \text{PAR}_{\text{males}} = \frac{0.2 \cdot 1.5}{0.2 \cdot 1.5 + 1.0} = 0.23 \]
\[ \text{PAR}_{\text{females}} = \frac{0.1 \cdot 3.0}{0.1 \cdot 3.0 + 1.0} = 0.23 \]

Although the AR is higher in females than in males, the PAR's are the same because males are exposed more frequently than females.
Summary

- Epidemiology is the science that studies the distribution of diseases and of their causal factors in populations.
- The major morbidity measures used in epidemiology are the prevalence, i.e. the disease frequency, and the incidence, i.e. the occurrence rate of new cases.
- Epidemiological studies can be either interventional or observational. In terms of their timing, studies can be of prospective or retrospective design.
- The effect of an exposure upon disease risk can be measured by the relative risk or the odds ratio.
- (Observational, retrospective) case-control studies do not allow the estimation of relative risks, only of odds ratios.
The aim of standardisation techniques is to remove, as far as possible, the effects of demographic or other differences when comparing the disease patterns observed in different populations. A standardised incidence rate thus equals the rate at which new cases are expected to arise overall if the population of interest had the same structure, with respect to all critical variables of interest, as a given standard population.

An important step in applying standardisation is the selection of the standard population since the actual value obtained after standardisation depends upon this choice. To a certain extent, however, the standard population can be chosen arbitrarily since the numerical outcome of the standardisation itself has no practical significance. It merely serves for the comparison of groups, not as a measure of absolute magnitude. The standard population may therefore either come from the study populations themselves (e.g. an average) or may be a population without any relation to the data under study. In general, however, the standard population should not be radically different from the populations to be compared. One example of a suitable standard population is the so-called ‘world standard population’, regularly defined by the WHO Cancer Research Agency in Lyon, France.

The formula for the standardised incidence rate $\gamma_S$ is:

$$\gamma_S = \frac{\sum_{i=1}^{k} s_i \cdot \gamma_i}{\sum_{i=1}^{k} s_i}$$

is called the 'standardised incidence rate'.

Appendix

Standardisation

Let a population be stratified into $k$ strata (e.g. age, gender) with incidence rates $\gamma_1, \ldots, \gamma_k$.

Let $s_1, \ldots, s_k$ be "standard" person-times, e.g. in a reference population.
In the present example, males are obviously at a higher risk of disease than females, and this sex difference is more pronounced in population 2 than in population 1. If the sex difference is ignored, the second population would nevertheless seem to face a disease risk that is approximately only half that of the first population.

However, a closer look at the sex ratios reveals that ignoring the sex difference meant comparing apples and pears. Whilst the first sample comprised twice as many observational units (i.e. person years) in males than in females, the sex ratio was one to five in the second sample. The reason why the second population appeared to be better off in terms of risk was therefore due to the fact that the corresponding sample was enriched with low-risk individuals.
Standardisation reveals that, contrary to the result obtained when ignoring the sex difference in terms of disease risk, the first population seems better off than the second one. It is interesting to note that this would be the case as long as the sex ratio (females:males) in the standard population is smaller than 2:1. For a sex ratio equal to 2:1, both standardised incidence rates are equal to 0.027. If the sex ratio in the standard population is larger than 2:1, then the second population would appear to face a lower overall risk since the excessively high risk of males would be compensated for by the low female risk.
In a situation where two sub-populations, say males and females, with different relative disease risks under exposure are considered as one, and this sub-structuring is ignored, the relative risk estimated from the mixed population will lie between the relative risks pertaining to the two sub-populations. However, this is only true as long as the exposure probabilities in the two sub-populations are identical, so that the admixture proportions are the same among exposed and non-exposed individuals. When differences in terms of the exposure probability exist and are ignored, the estimation of relative risks from mixed populations can be misleading. If, in the present example, all exposed individuals were male and all non-exposed were female, than the ensuing relative risk of $0.50/0.02=25$ would represent a gross over-rating of the true relative risk.

Differences in exposure probability can be balanced during the sampling process. However, such balancing would be problematic if the stratification of the population into subpopulations with different relative risks is not obvious to the researcher.
With odds ratios, ignoring differences between sub-populations in terms of disease risk can have serious consequences, even if the exposure probabilities are the same in all sub-populations. There is no way of telling whether the odds ratio in the joint population is larger, or smaller, or somewhere in between the odds ratios of individual sub-populations. In the present example, lumping males and females together would potentially lead to an underestimation of the true association, particularly if the sex ratio in the joint sample is around 1:2.