4 Epidemiology

Video4.1 - Epidemiological study designs

4.1 Epidemiological study designs

According to the British Medical Journal (BMJ), Epidemiology is defined as follows.

> Epidemiology is the study of how often diseases occur in different groups of people and why.

Thus, epidemiological studies compare differences in disease between the following groups.

- **Individual** - The base unit is a person. Data about individuals are recorded, such as health status, possible exposures of interest (e.g. smoking), and other important covariates (e.g. age, sex, etc.) The individual level impact of the exposure on disease risk is assessed after adjusting for the other covariates.
- **Spatial ecological** - The base unit is a group of individuals living within a small geographical unit such as an electoral ward. Data about the total level of disease, average levels of exposure (e.g. percentage of people that smoke) and other important covariates (e.g. average age) are collected for each geographical unit. The population level impact of disease is assessed by a comparison between the populations living in the different spatial units.
- **Temporal ecological** - The base unit is a time interval, such as a day or year, and data are collected about the same population for numerous consecutive time intervals. For example, the number of disease cases per day from the population are collected, together with a measure of exposure (e.g. average air pollution concentrations) and other covariates such as average temperature. The impact of disease is then assessed by a comparison in disease rates between the same population in the different time intervals.

Note The first of these studies is an individual level study, from which **individual level cause and effect can be established**. The latter two are population or ecological level studies, which compare disease risk between groups of individuals rather than between individuals. The latter is much weaker, and thus ecological studies **cannot be used to estimate individual level cause and effect**.

How do you know that the people in a group with the disease are the same ones in that group that had the exposure? The generic term for wrongly assuming that an estimated group level relationship holds at the individual level is the **ecological fallacy**, and the difference between the group and individual level effects is known as **ecological bias**.

Example Figure 18 shows the relationship between death rate due to breast cancer and dietary fat intake, where the unit is a country. A clear linear relationship is observed, with increasing fat intake being associated with higher death rates.

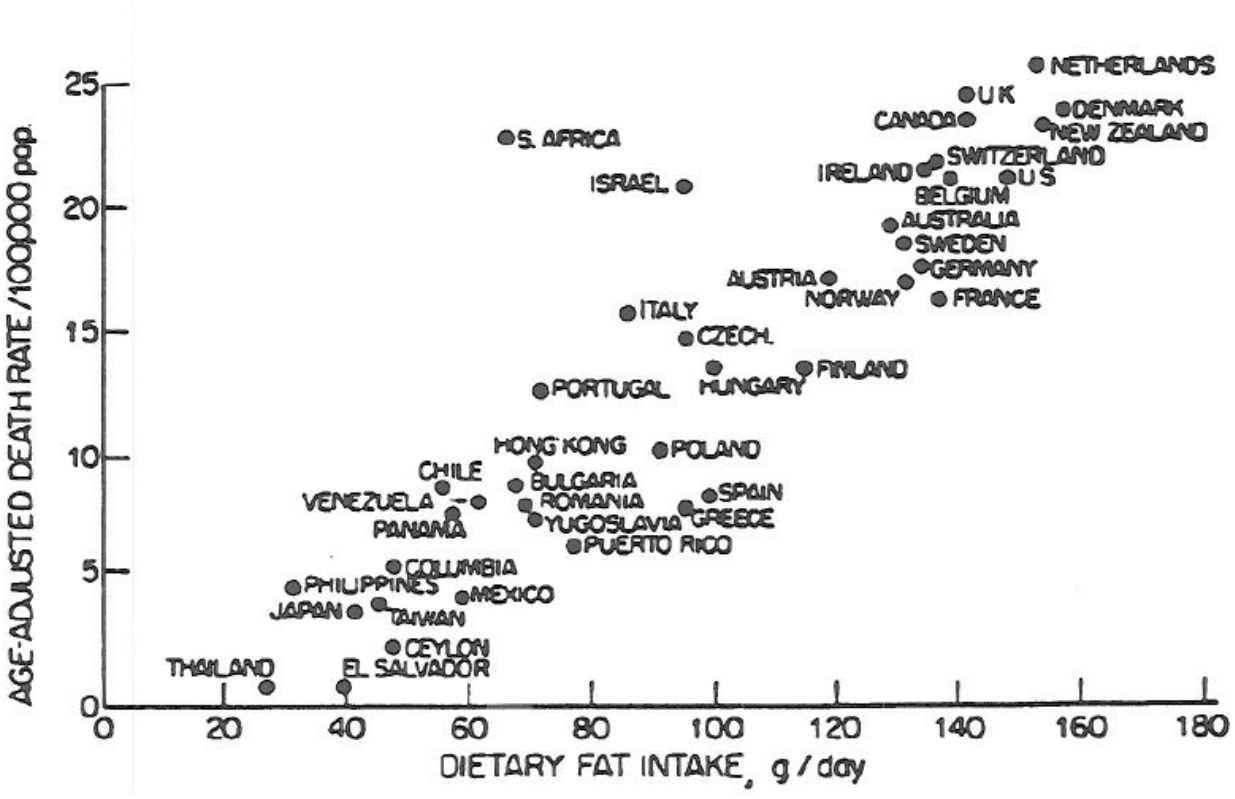

Figure 18: Age-adjusted death rate per 100,000 people is plotted against fat intake in g/day. The plot shows a strong positive linear relationship between the two variables.

How do we know that the women with high fat intake were the same ones that died from breast cancer. Assuming they are is the ecological fallacy.

Additionally, dietary fat intake is only one possible explanation, if the populations living in these countries differ in some other way then that might be causing the observed association. Examples here might be different underlying genetic susceptibility, breastfeeding rates (protective), physical activity levels, etc. These other factors are an example of confounders, which we discuss later in the course. Within individual level studies there are two types of study designs.

Definition In a **cohort study**, a group of individuals is selected and are followed over time, possibly for many years. During the follow up period data are collected periodically about the development of disease and other factors of interest such as age, sex, smoking status, etc. To study associations with the development of disease, subjects are classified on the basis of their characteristics before its development, such as smokers vs. non-smokers. These groups are compared in terms of their subsequent mortality or morbidity rates. This study works from **cause to effect**, since baseline data on risk factors are collected before disease onset for these individuals (possible causes), and then are related to subsequent disease occurrence (effect).

- **Advantages**
  - Provides a complete description of health after exposure.
  - Allows the study of multiple potential health effects of a given exposure, including benefits as well as risks.
  - Allows the calculation of rates of disease in exposed and unexposed individuals.
  - Gives a flexibility in choosing which variables are recorded.

- **Disadvantages**
  - Large number of subjects are required to study a rare disease.
  - Potential long duration for follow-up, thereby making the study lengthy and hence expensive to conduct.
  - Maintaining contact with all cohort members is difficult, thus you lose data.

Definition In a **case-control study**, individuals with a particular disease (the cases) are selected for comparison with a series of individuals in whom the disease or condition is absent (the controls). Cases and controls are compared with respect to existing or past exposures thought to be relevant to the development of the disease or condition under study. Thus, this study works from **effect to cause**, because the patients are identified based on whether or not they have the disease (effect), and then possible causes are asked about via interview.

- **Advantages**
  - Well suited to studying rare diseases or diseases with long latency.
  - Relatively quick and inexpensive to conduct.
  - Comparatively few subjects are required compared to a cohort study.
  - It allows multiple potential causes of disease to be studied.
- **Disadvantages**
  - Relies on patient recall for information on past exposures which will likely be inaccurate.
  - Control for confounding variables might be incomplete since they might not be available.
  - The rates of disease in exposed and unexposed individuals cannot be determined.

Definition **Causation** of an exposure or treatment on a disease outcome can be estimated from an experimental study such as a clinical trial. However, in general an observational study cannot claim to prove causation, due to the number of factors that are not controlled for. To have definitive evidence for causation from observational data you need to show the following:

- **strong**: a large effect size that is unlikely to be due to an alternative factor.
- **graded**: a dose-response relationship exists, that is the higher the exposure the worse the response.
- **independent**: the association persists after adjusting for all known confounders.
- **consistent**: the same association has been shown in many different studies.
- **reversible**: if you remove the exposure the risk of disease reduces again.
- **plausible**: is there a biologically plausible reason why the exposure would cause disease?
- **in temporal sequence**: is the exposure present before the disease onset.

Example During the first half of this century, routinely collected mortality data established that deaths from lung cancer were increasing at an alarming rate. In England and Wales in 1922 there were 612 deaths from lung cancer, which had risen by a factor of 15 to 9287 by 1947. A number of possible explanations were mooted.

- The population was larger and older (on average) in 1947 than in 1922.

- Improved diagnosis. This is almost certainly responsible for some of the increase. However, this is unlikely to be responsible for such a large increase, especially since increases in urban areas (with good diagnostic facilities in hospitals) and rural areas (with poorer diagnostic facilities) were similar.
- Increasing tobacco consumption, as evidenced by Figure 19.

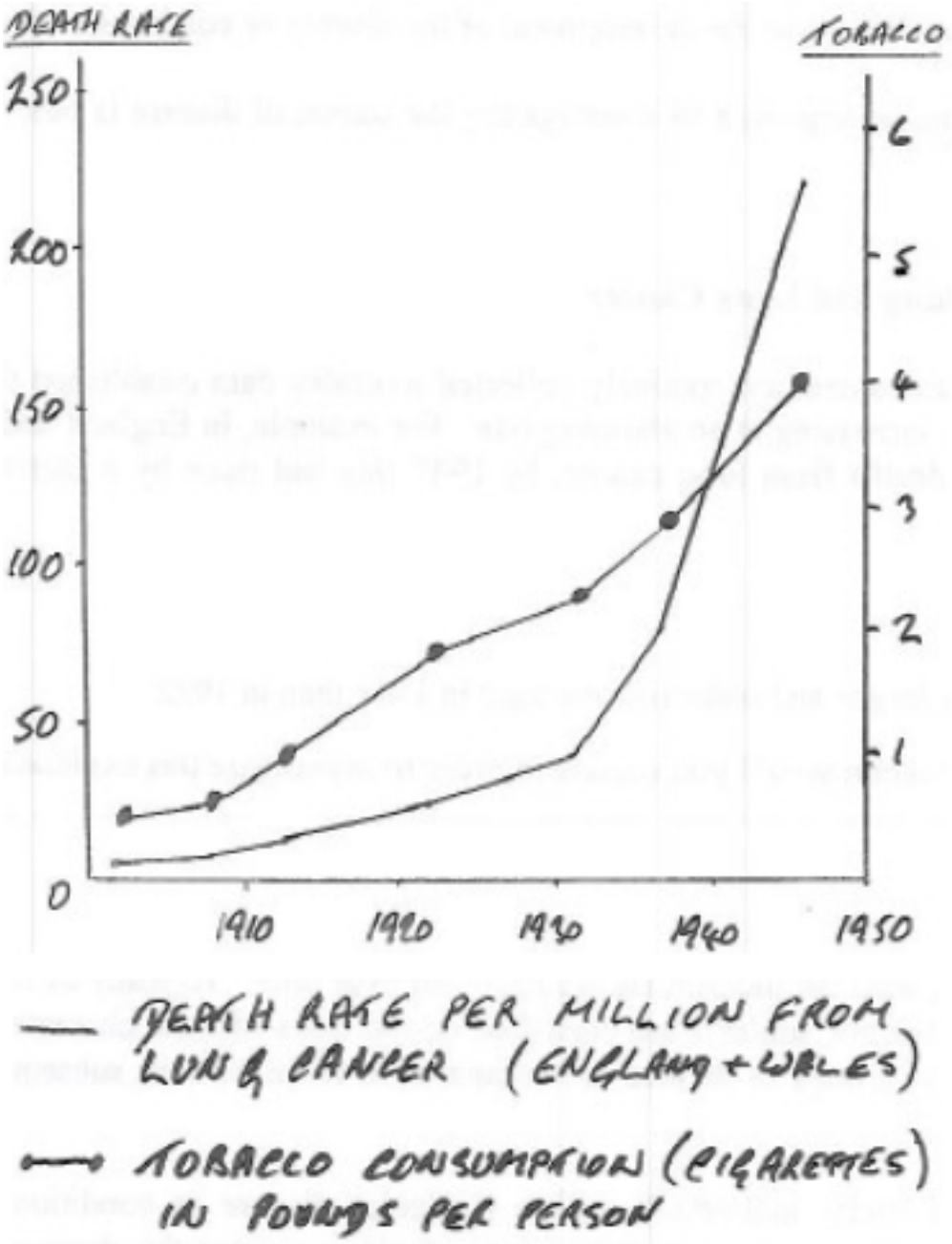

Figure 19: The death rate from lung cancer per one million people (y-axis on left) and the tobacco consumption in pounds per person (y-axis on right) are plotted against time. The two y-axes are on different scales and the plot displays a clear strong positive relationship.

- This was strongly suspected by many people as the main cause. The graph above shows the increasing

trends in lung cancer mortality and tobacco consumption. Also, many physicians were aware that the overwhelming majority of their lung cancer patients were smokers.

This is as much as can be achieved with routinely collected data, causation cannot be proved yet. The next step is to conduct an observational study on individual subjects. The study design of first choice is a case - control study, since an answer is required quickly. A cohort study would involve following up many subjects over a long period of time. We return to this example in Measuring the association between a risk factor and disease.

Example - Chilli peppers increase life expectancy Claiming causation based on a single preliminary study results in newspaper articles making outlandish claims such as this one from the *Independent*: People who regularly eat chilli peppers live longer, research suggests. The *daily express* titled the same news story as How to live longer: Eat this more than four times a week to lower risk of early death. Those stories are based on research published in the Journal of the American College of Cardiology; so can eating chillies prolong your life? The paper quotes several hazard ratios that relate to death due to any cause, various heart diseases and cerebrovascular disease. All hazard ratios are significantly less than 1. The only conclusion that can be drawn from those results are that there appers to be a significant association between eating chillies and life expectancy. To establish any causal link requires more studies that show that all 7 bullet points above are met. The newspaper articles appear to sell this preliminary research as causal evidence which could not be further from the truth; the truth however rarely makes for such interesting headlines as the two above.

4.2 Sensitivity and Specificity

Video4.2 - Sensitivity and Specificity+Measuring disease rates

Definition Consider a screening test that is given to people who are suspected of having a particular disease. Most screening tests are not 100% accurate, and thus their quality can be measured by two metrics:

$$\textbf{Sensitivity} = \frac{\text{Number of diseased people who screen positive}}{\text{Total number of diseased people}}.$$

$$\textbf{Specificity} = \frac{\text{Number of healthy people who screen negative}}{\text{Total number of healthy people}}.$$

The negative of these statistics are:

$$\textbf{False negative rate (FNR)} = \frac{\text{Number of diseased people who screen negative}}{\text{Total number of diseased people}}.$$

$$\textbf{False positive rate (FPR)} = \frac{\text{Number of healthy people who screen positive}}{\text{Total number of healthy people}}.$$

Note, that:

$$\text{FNR} = 1 - \text{sensitivity};$$

$$\text{FPR} = 1 - \text{specificity}.$$

Example Consider a new screening test for bowel cancer, which is given to 1000 people, and the results are displayed in Table 19 below.

Table 19: 2x2 table summarising the risk of bowel cancer and the results of a new screening test.

	Disease		**Total**
	Yes	**No**	
Positive test	45	25	**70**
Negative test	5	925	**930**
Total	**50**	**950**	**1000**

Then

$$\textbf{Sensitivity} = \frac{45}{50} = 0.90 \quad \textbf{FNR} = \frac{5}{50} = 0.10.$$

$$\textbf{Specificity} = \frac{925}{950} = 0.974 \quad \textbf{FPR} = \frac{25}{950} = 0.026.$$

4.3 Measuring disease rates

Definition There are 3 common crude measures of disease:

- **Prevalence rate** $= \dfrac{\text{number of cases of disease at a specified time point}}{\text{population size}}$.
- **Incidence rate** $= \dfrac{\text{number of new cases of disease during a specified time period}}{\text{total population at risk (free of disease at start of period)}}$.
- **Mortality rate** $= \dfrac{\text{number of deaths in a specified time period}}{\text{midperiod population}}$.

These quantities are usually multiplied by 100,000, so as to get the prevalence, incidence or mortality rate per 100,000 people. The time period for incidence and mortality is usually a year.

Example

- A prevalence of 145 asthma cases per 100,000 means that in a population of 100,000 people there are 145 people with asthma.
- An incidence of 145 new asthma cases per 100,000 per year means that in a population of 100,000 people there were 145 new people diagnosed with asthma in the year.
- A mortality rate of 145 per 100,000 for asthma per year means that in a year 145 people died of asthma from a population of 100,000.

Notes Mortality data come from death records, incidences come from hospital records such as cancer registries, while prevalence data are much harder to get (an exception is cancer registries), but could be estimated from surveys such as the Scottish health Survey Scottish health Survey. You also need data on the denominators in the above rates on population sizes, which can be obtained from the census that is carried out in the UK every 10 years (1991, 2001, 2011, etc).

Example Data on the number of admissions to hospital due to respiratory diseases (incidence) in 2011 were collected from the Scottish Statistics database, for the 271 Intermediate geographies that make up the Greater Glasgow and Clyde health board. The observed number of admissions are shown in Figure 20 below.

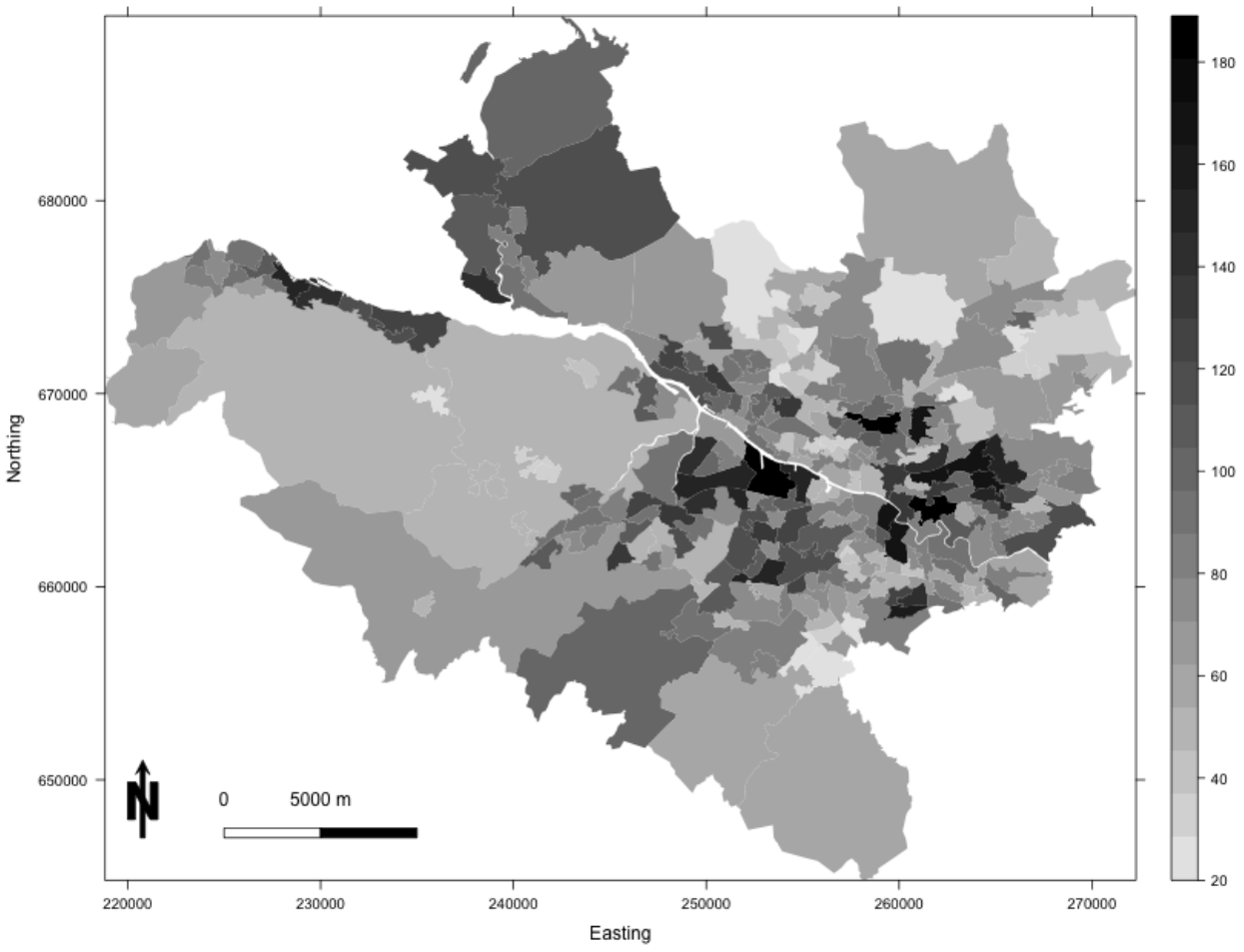

Figure 20: The number of admissions to hospital due to respiratory diseases is plotted for every intermediate geography in the GGCHB. Darker colours relate to more hospital admissions and lighter colors to fewer admission.

A few high incidence areas stand out. However, each area has different population sizes and demographics (e.g. age and sex) structures, and one would naturally expect areas with larger and older populations to exhibit a higher incidence.

Example The number of lung cancer deaths in Scottish males in 1990 and 2005 are summarised in Table 20 below.

Table 20: The number of deaths and population estimates in Scotland for the years 1990 and 2005.

Year	No. deaths	Population
1990	2,670	2,443,865
2005	2,195	2,456,109

The raw or crude mortality rate (per 100,000) for both years is:

- **1990** - Crude mortality rate $= \frac{2670}{2{,}443{,}865} \times 100{,}000 = 109.3$ per 100,000.
- **2005** - Crude mortality rate $= \frac{2195}{2{,}456{,}109} \times 100{,}000 = 89.4$ per 100,000.

From 1990 to 2005 the crude mortality rate fell from 109.3 to 89.4 per 100,000, which is a reduction of:

$$\frac{109.3 - 89.4}{109.3} \times 100 = 18.2\%.$$

Is this a valid comparison?

Table 21 below shows the mortality rates for 1990 and 2005 separately within each five year age group for males, along with the percentage change for each group. Does an 18% reduction seem valid now?

Table 21: Mortality rates for lung cancer among Scottish males per 100,000 is given for the years 1990 and 2005. The data are presented in 5 year age bands and the % reduction of cancer mortality rates from 1990 to 2005 are given as well.

Age group	Mortality rate 1990 (per 100,000)	Mortality rate 2005 (per 100,000)	% reduction from 1990 to 2005
Under 5	0	0	0
5-9	0	0	0
10-14	0	0	0
15-19	0	0	0
20-24	0	0	0
25-29	0	0	0
30-34	1.1	0	100
35-39	1.2	1.6	-33
40-44	10.3	8.2	20
45-49	29.7	21.1	29
50-54	74.2	50.3	32
55-59	160.9	93.7	42
60-64	323.4	196.1	39
65-69	470.7	299.6	36
70-74	628.4	432.3	31
75-79	787.1	612.0	22
80-84	901.3	714.1	21
85+	801.8	655.3	18
All ages	**109.3**	**89.4**	**18**

As you can see 18% seems like a vast underestimate in the mortality rate when considering the individual age groups. Ignoring the under 40 age groups which have almost no mortalities, the reductions range between 18% and 42%, while the overall crude figure is 18%. How has this happened?

The answer is that between 1990 and 2005 the distribution of the age structure of the Scottish population has changed, and in particular got older. For example, in 1990 12.1% and 1.9% of the population were over 65 and 80 respectively, which had risen to 14.1% and 2.8% by 2005. Thus as the population were older in 2005 compared to 1990, then they are likely to have more mortalities as the risk for older age groups is larger. This results in the crude rates being biased and not comparable. Therefore, as a general (and very important) rule: **Crude rates should not be used to compare populations**.

4.4 Standardisation

Video4.3 - Direct standardisation

Adjusting for population demographics is known as standardisation, which has two general flavours, direct and indirect. The discussion that follows is based on mortality, but the same approaches are used for incidence and prevalence.

4.4.1 Direct standardisation

Direct standardisation is used to compare the rate of disease in 2 separate populations with different demographics. It requires a standard or reference population, and the rates of disease for two separate study populations are transferred to this reference population to make them directly comparable.

Definition The most common standard populations are the **European standard** population and the **World standard** population (Table 22), both of which contain 100,000 people. These are hypothetical populations with an age structure that approximates the average for Europe and the World respectively.