
MAPS | eCademy
Webinar Series
Part 1 of Webinar Series:
Clinical Statistics
for Non-Statisticians

Introduction

Firas Dabbous, MS, PhD

Associate Director of HEOR

Abbvie

Disclaimer

The views expressed in this Webinar are those of the presenters, and are not an official position statement by MAPS, nor do they necessarily represent the views of the MAPS organization or its members.

Objectives

- To define some terminologies in statistics
- To explain different types of data (quantitative vs. qualitative)
- To describe different ways in presenting summary data (such as proportions, means, medians, range and standard deviation)
- To explain 95% confidence intervals
- Bivariate analyses and appropriate statistical testing for quantitative and qualitative variables

Biostatistics

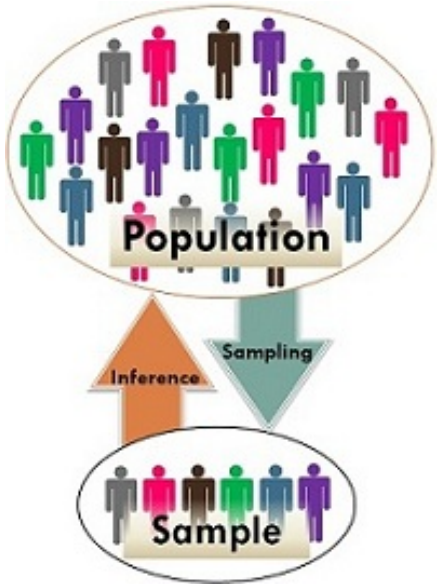
- Defined by Merriam-Webster as "the statistical processes and methods applied to the collection, analysis, and interpretation of data and especially data relating to human biology, health, and medicine"
- ***Statistical analyses*** are used to manipulate, summarize, and investigate data, so that useful decision-making information are generated

Types of Statistics

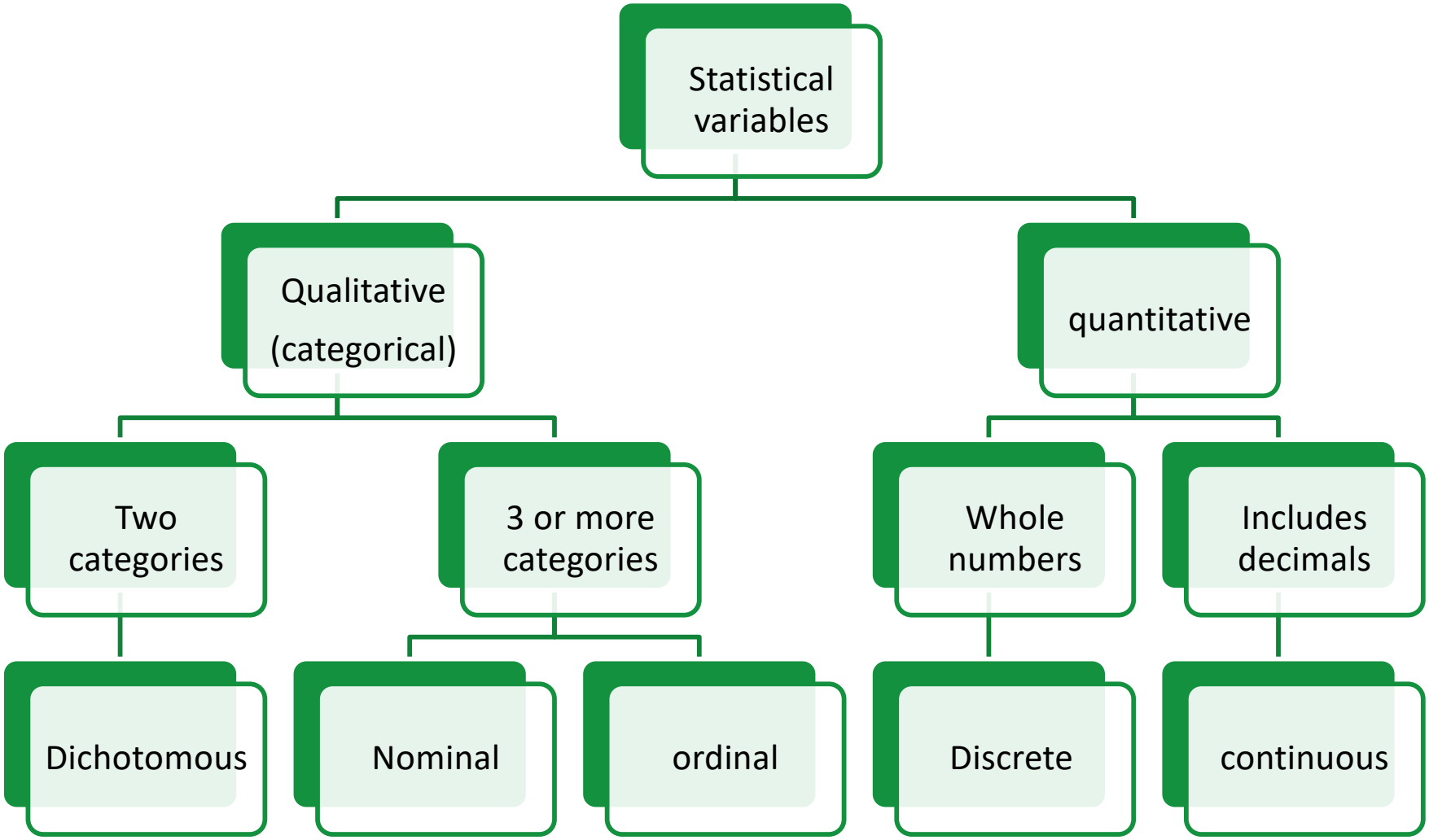
- *Descriptive statistics* are used to describe the population of interest
  - Summary outcomes include proportion, mean, standard deviation, median, etc..
- *Inferential statistics* are used to make general conclusions using study samples about populations of interest
  - Because it is often too expensive and not-feasible to study the general population, statistical method are used to make inferences on the larger population.
  - Samples are usually part of the large population and sample statistics are imperfect representations of the corresponding population.

Population and Sample

- Population is the entire group of patients
  - Example, a researcher is interested in studying smoking (variable 1) and lung cancer (variable 2) in Medicare patients
- Sample is representative of the larger population
  - Should have similar characteristics as the population it is representing



Variable types



Qualitative (categorical) variables

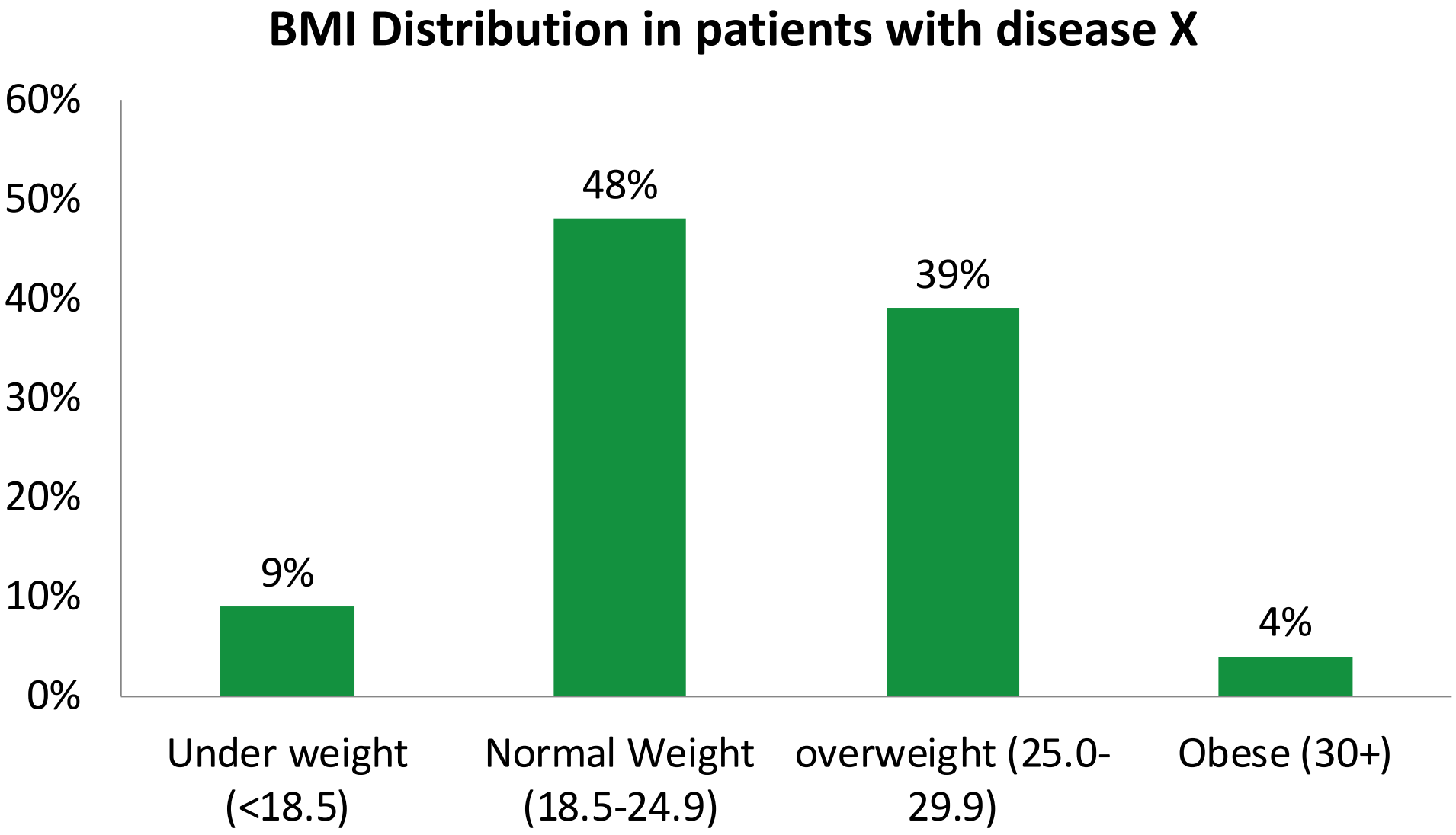
- Measures non-numeric qualities and take category or label value:
  - Dichotomous or binary variables, those that have only two categories
    - E.g. Presence of a certain disease (yes, no)
  - Nominal variables have three or more mutually exclusive categories without an apparent order
    - E.g. Hair color, race/ethnicity, blood type
  - Ordinal variables have three or more categories and are ordered such that values in one category are larger or smaller that those in another
    - E.g. visual analog Scale, Cancer staging

Proportions

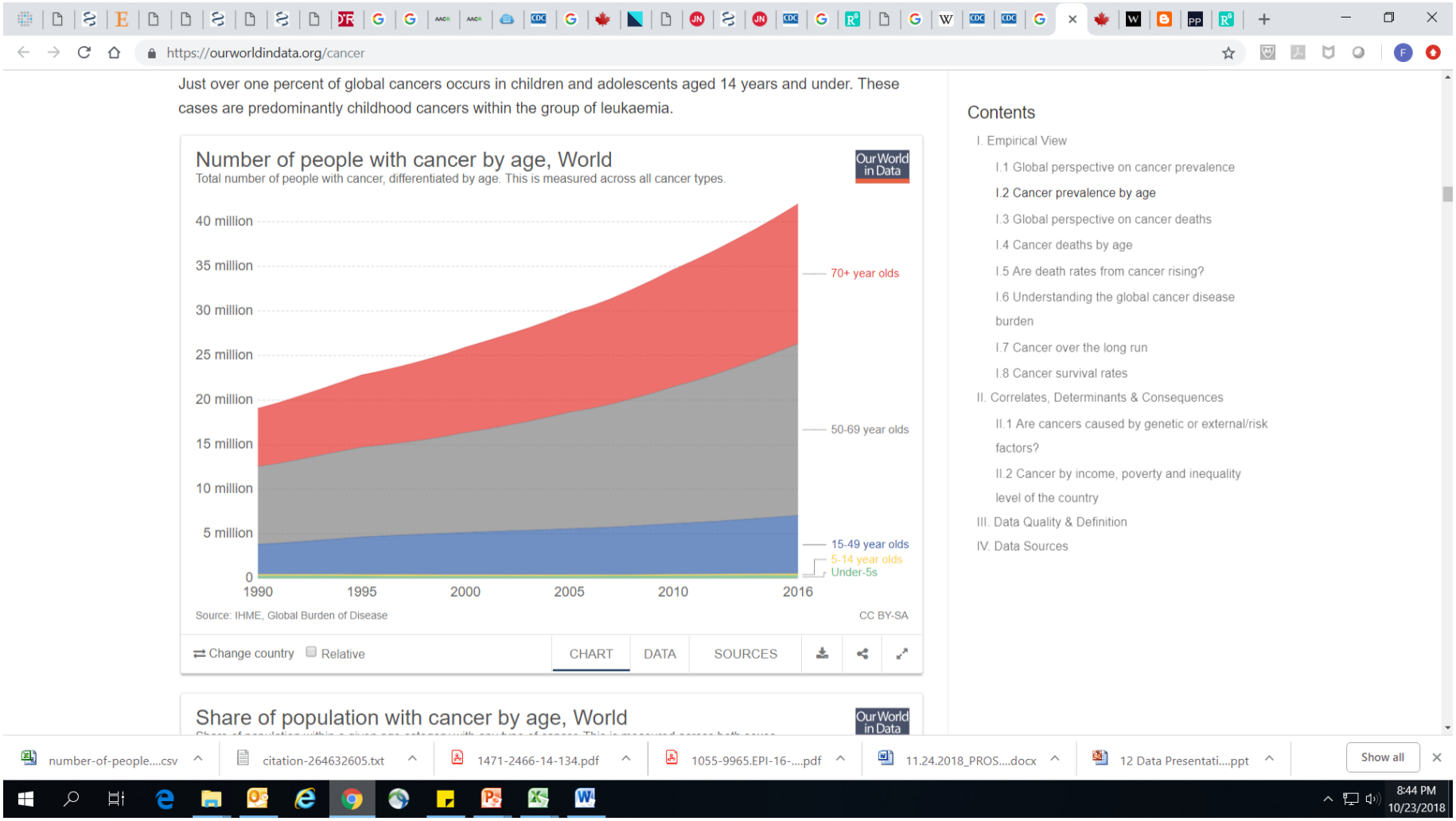
- Proportion is the number of patients with a certain characteristic of interest divided by the total sample of patients (in other words prevalence)

$$\frac{Number\ of\ obsese\ patients\ in\ the\ sample}{total\ number\ of\ patients\ in\ teh\ sample}$$

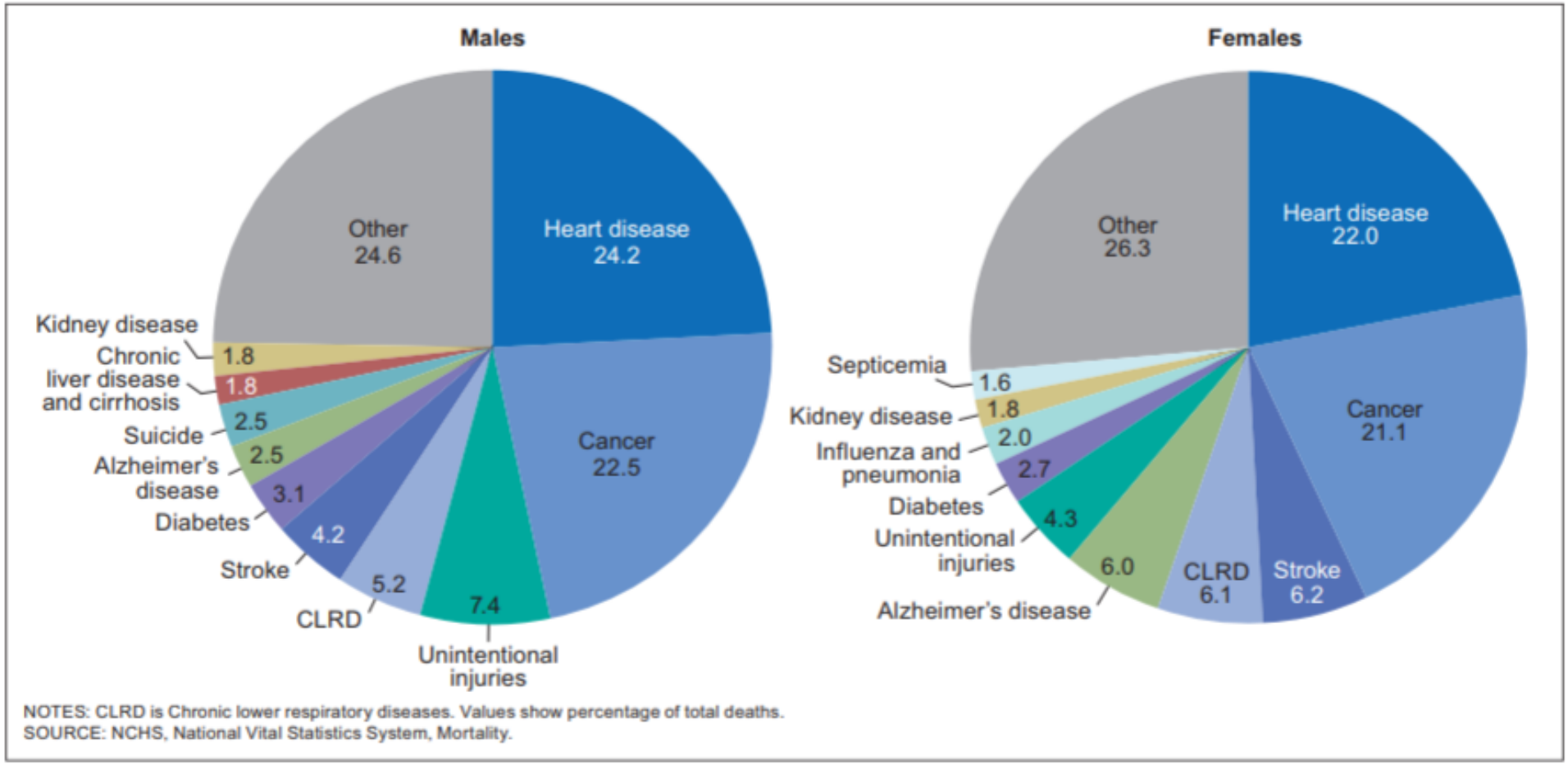
Bar Chart


BMI Distribution in patients with disease X
60%
50%
40%
30%
20%
10%
0%
9%
48%
39%
4%
Under weight (<18.5)
Normal Weight (18.5-24.9)
overweight (25.0-29.9)
Obese (30+)

Area Chart


https://ourworldindata.org/cancer
Just over one percent of global cancers occurs in children and adolescents aged 14 years and under. These cases are predominantly childhood cancers within the group of leukaemia.
Number of people with cancer by age, World
Total number of people with cancer, differentiated by age. This is measured across all cancer types.
Our World in Data
40 million
35 million
30 million
25 million
20 million
15 million
10 million
5 million
0
1990
1995
2000
2005
2010
2016
70+ year olds
50-69 year olds
15-49 year olds
5-14 year olds
Under-5s
Source: IHME, Global Burden of Disease
CC BY-SA
Change country
Relative
CHART
DATA
SOURCES
Share of population with cancer by age, World
Our World in Data
Contents
I. Empirical View
I.1 Global perspective on cancer prevalence
I.2 Cancer prevalence by age
I.3 Global perspective on cancer deaths
I.4 Cancer deaths by age
I.5 Are death rates from cancer rising?
I.6 Understanding the global cancer disease burden
I.7 Cancer over the long run
I.8 Cancer survival rates
II. Correlates, Determinants & Consequences
II.1 Are cancers caused by genetic or external/risk factors?
II.2 Cancer by income, poverty and inequality level of the country
III. Data Quality & Definition
IV. Data Sources
number-of-people....csv
citation-264632605.txt
1471-2466-14-134.pdf
1055-9965.EPI-16-....pdf
11.24.2018_PROS....docx
12 Data Presentati....ppt
Show all
8:44 PM
10/23/2018

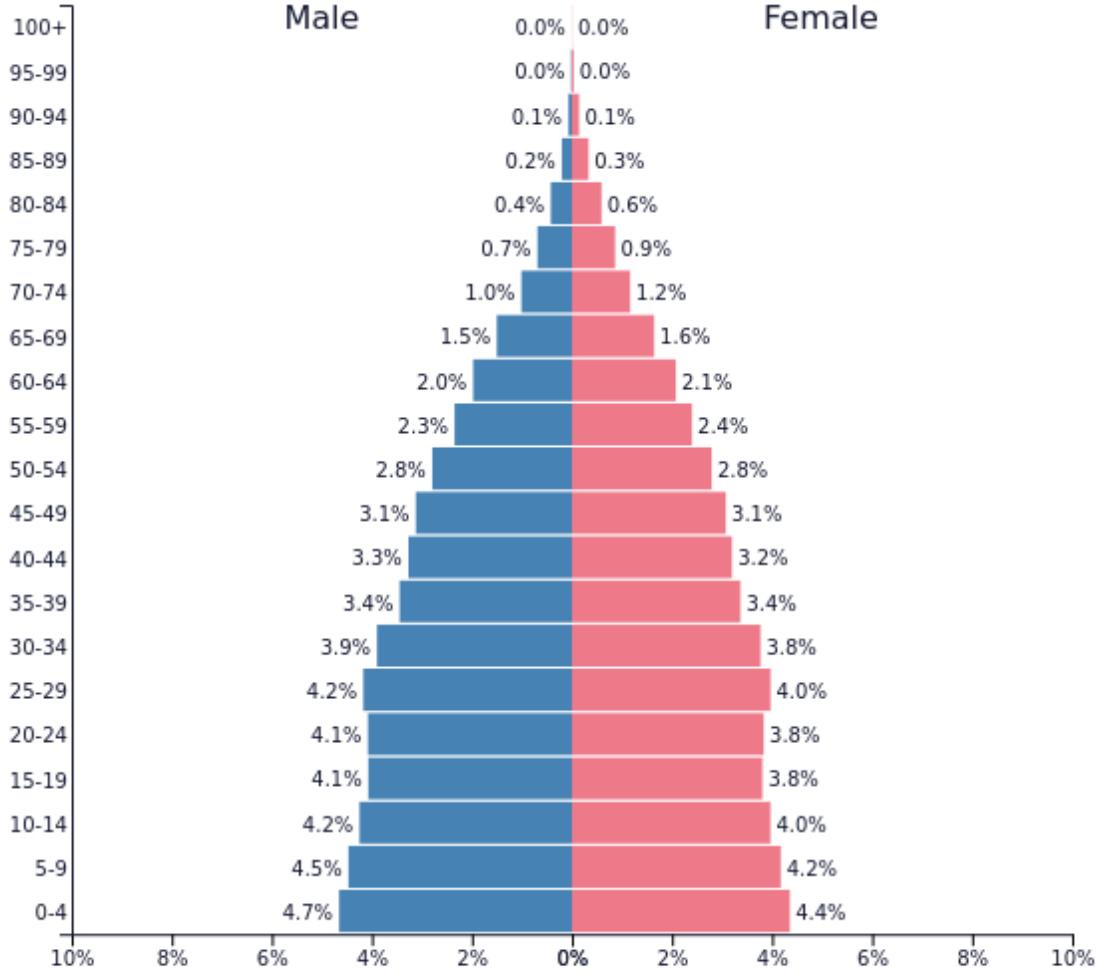
Pie Charts



NOTES: CLRD is Chronic lower respiratory diseases. Values show percentage of total deaths.
SOURCE: NCHS, National Vital Statistics System, Mortality.

Figure 1. Percent distribution of the 10 leading causes of death, by sex: United States, 2016

Population pyramid



Ratio

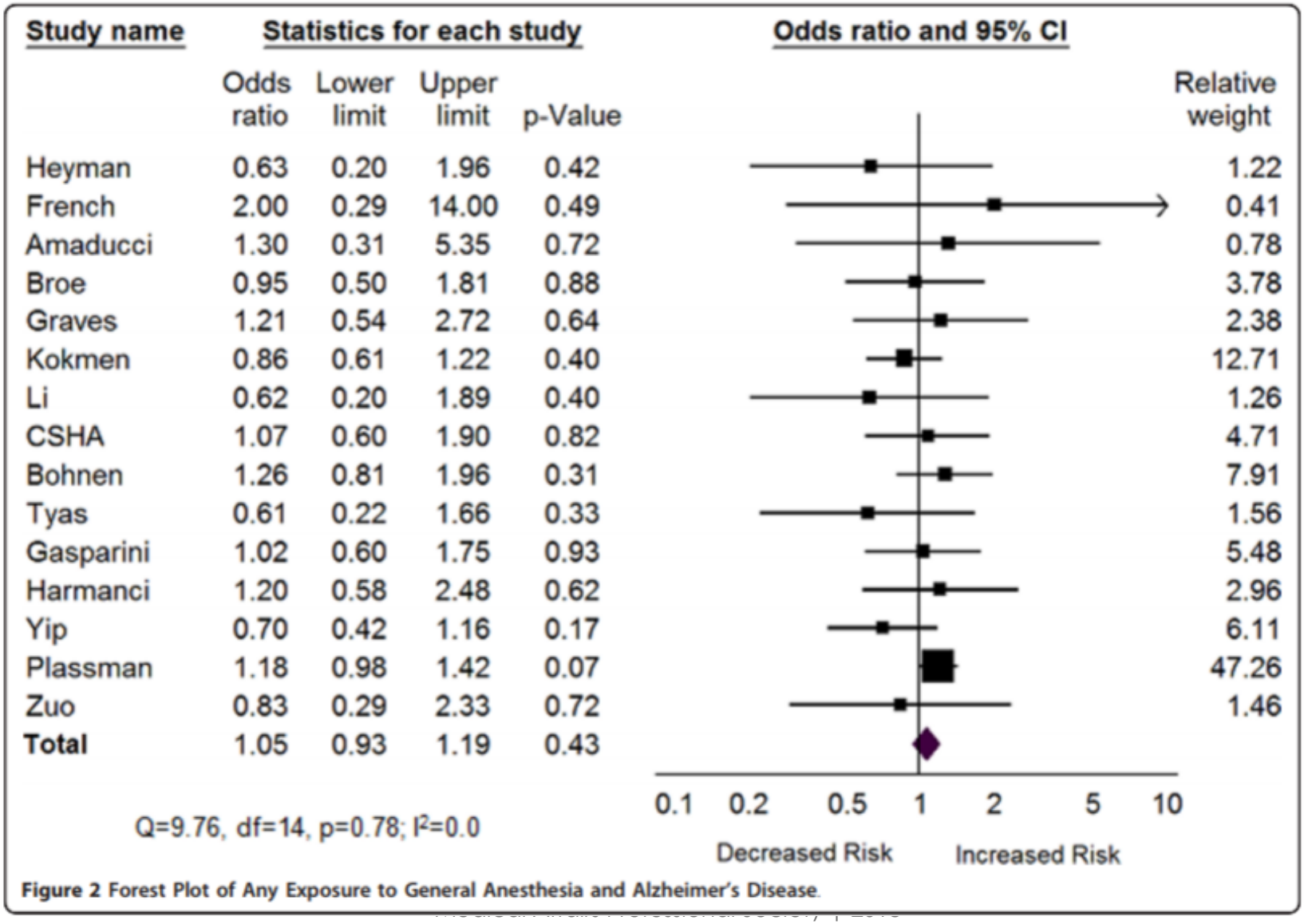
- Is the relative magnitude between two quantities
  - E.g. Number of obese in males/ number of obese in females
  - E.g number of cases to controls or in an analytical setting odds ratios, risk ratios or rate ratios (to be defined in later webinars)

Presenting Ratios in table format

User Characteristics	Bivariate		Multivariate Model	
	UOR (95% CI)	*p*-Value	AOR (95%CI)	*p*-Value
Gender				
Male	1 (Reference)		1 (Reference)	
Female	1.07 (0.72–1.65	0.73	1.00 (0.64–1.58)	0.10
Age group				
<30	1 (Reference)		1 (Reference)	
30–39	2.00 (1.21–4.21)	0.07	2.06 (0.93–4.57)	0.08
40–49	2.20 (1.25–4.29)	0.03 *	2.20 (1.01–4.81)	0.05
50–59	1.64 (0.95–2.62)	0.21	1.71 (0.74–3.94)	0.21
⩾60	2.47 (1.33–4.45)	0.05 *	3.19 (1.16–8.78)	0.03 *
Reason for HIV test				
Personal/both	1 (Reference)		1 (Reference)	
Medical indication	5.23 (2.59–7.91)	<0.01 *	4.84 (2.99–7.84)	<0.01 *
Status disclosure				
No	1 (Reference)		1 (Reference)	
Yes	1.10 (0.54–2.22)	0.79	0.77 (0.36–1.64)	0.50

* Statistically significant at 5% level.

Forest Plot



Study name	Odds ratio	Lower limit	Upper limit	p-Value	Relative weight
Heyman	0.63	0.20	1.96	0.42	1.22
French	2.00	0.29	14.00	0.49	0.41
Amaducci	1.30	0.31	5.35	0.72	0.78
Broe	0.95	0.50	1.81	0.88	3.78
Graves	1.21	0.54	2.72	0.64	2.38
Kokmen	0.86	0.61	1.22	0.40	12.71
Li	0.62	0.20	1.89	0.40	1.26
CSHA	1.07	0.60	1.90	0.82	4.71
Bohnen	1.26	0.81	1.96	0.31	7.91
Tyas	0.61	0.22	1.66	0.33	1.56
Gasparini	1.02	0.60	1.75	0.93	5.48
Harmanci	1.20	0.58	2.48	0.62	2.96
Yip	0.70	0.42	1.16	0.17	6.11
Plassman	1.18	0.98	1.42	0.07	47.26
Zuo	0.83	0.29	2.33	0.72	1.46
Total	1.05	0.93	1.19	0.43	

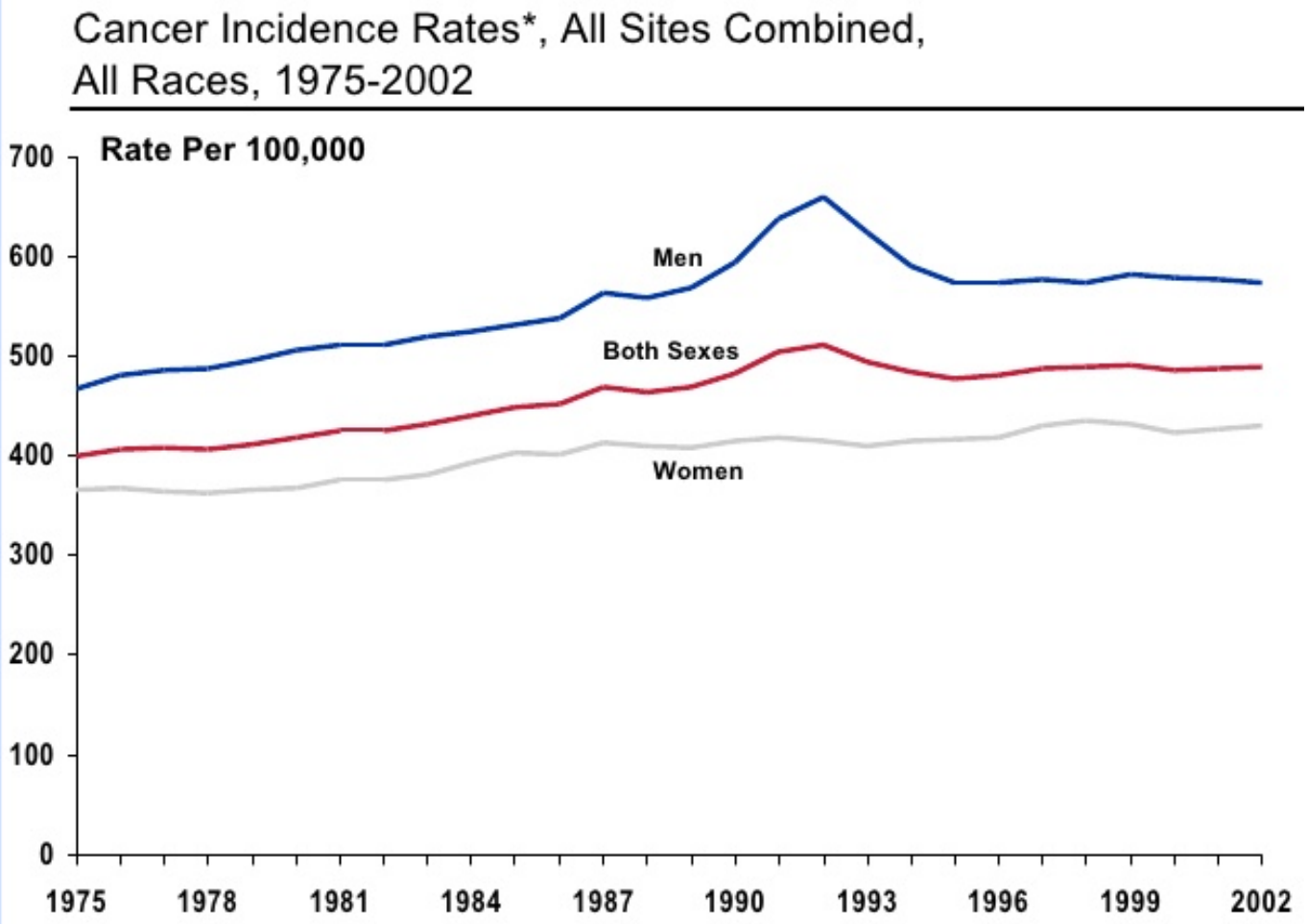
Q=9.76, df=14, p=0.78; I^2=0.0

Figure 2 Forest Plot of Any Exposure to General Anesthesia and Alzheimer's Disease.

Rate

- Is another measure of frequency of an event in a certain population over a certain period of time. (i.e. Incidence rate)

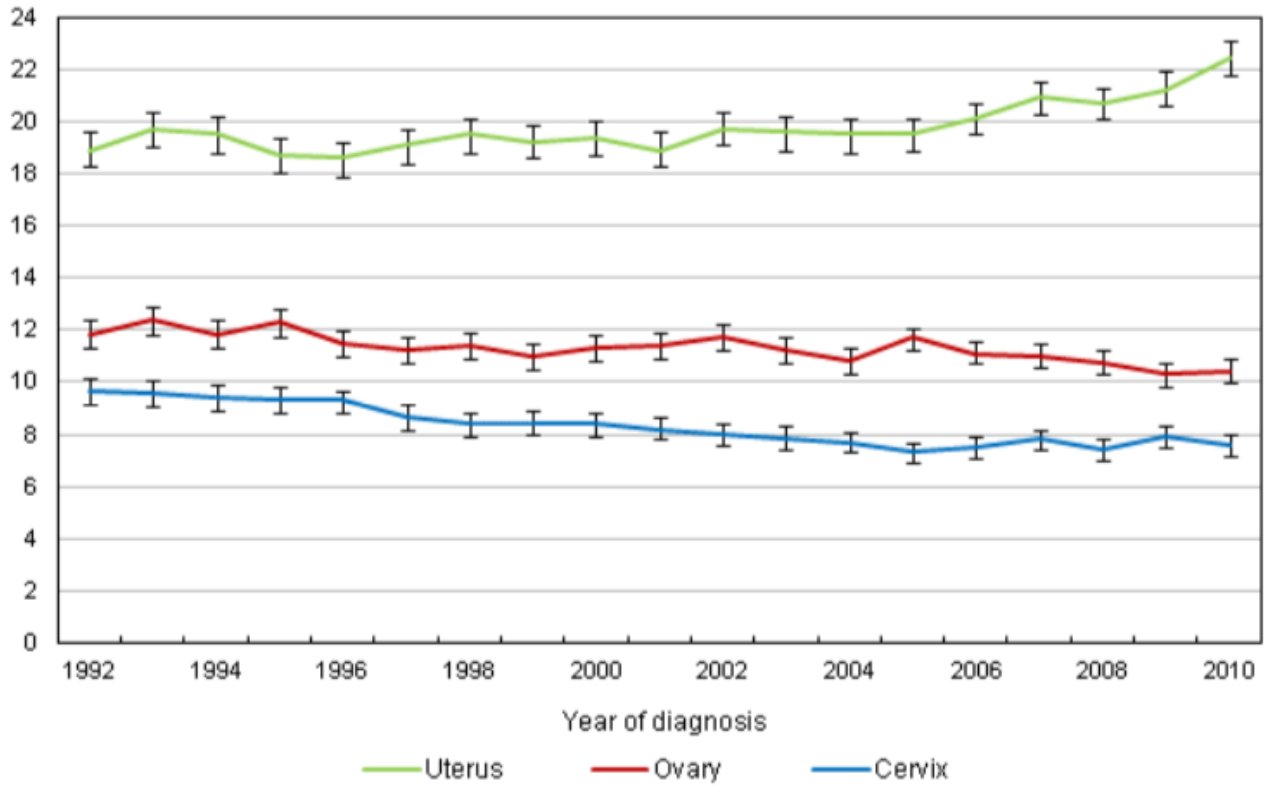
Line Graph



*Age-adjusted to the 2000 US standard population and adjusted for delay in reporting.
Source: Surveillance, Epidemiology, and End Results Program, 1973-2002, Division of Cancer Control and Population Sciences, National Cancer Institute, 2005.

Line Graph

Chart 1
Incidence rate, by type of cancer and year, age-standardized, Canada, 1992 to 2010

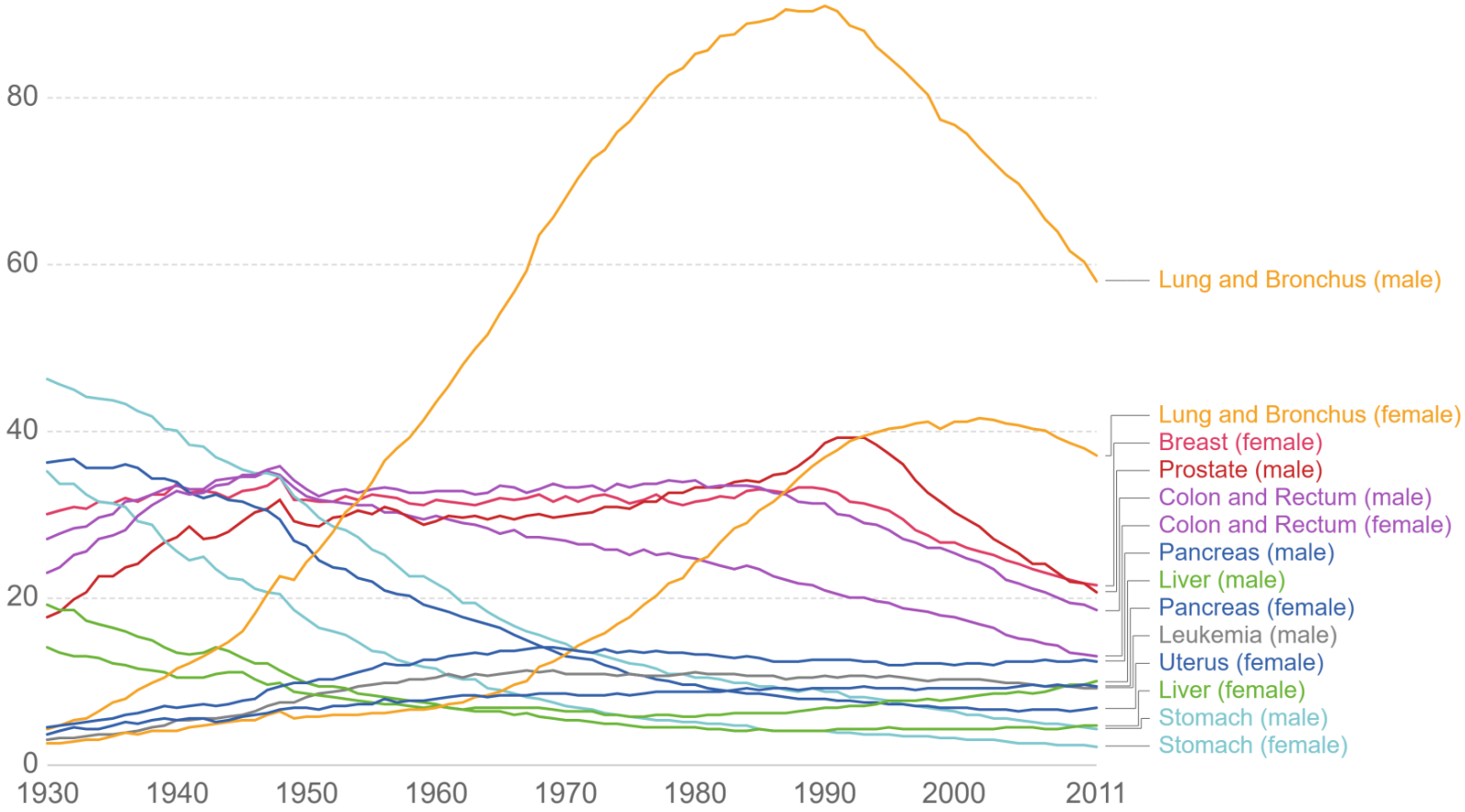


Note: The vertical error bars overlaid on the trend lines indicate the 95% confidence intervals. Confidence intervals indicate the degree of variability in the estimate and enable more valid comparisons of differences between estimates.
Source: Canadian Cancer Registry, Statistics Canada.

Time series

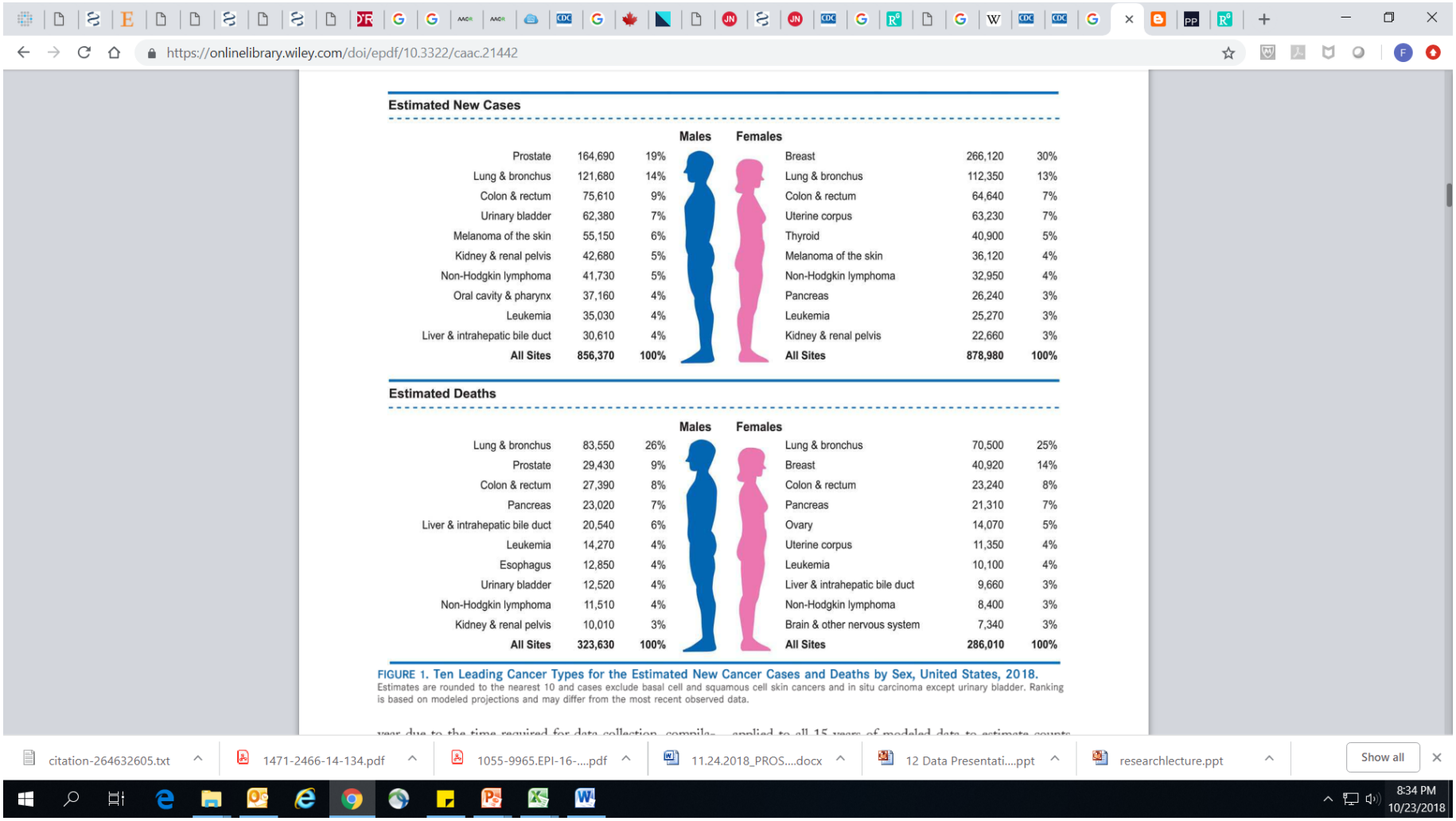
Cancer death rates in the United States over the long-run

Age-standardized death rates from various forms of cancer in males and females, measured as the number of deaths per 100,000 individuals. Age-standardization is based on normalisation to the standard US population structure in the year 2000.



Source: American Cancer Society

Incidence rates and mortality rates in the USA



https://onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21442

Estimated New Cases

	Males			Females	
Prostate	164,690	19%		Breast	266,120	30%
Lung & bronchus	121,680	14%		Lung & bronchus	112,350	13%
Colon & rectum	75,610	9%		Colon & rectum	64,640	7%
Urinary bladder	62,380	7%		Uterine corpus	63,230	7%
Melanoma of the skin	55,150	6%		Thyroid	40,900	5%
Kidney & renal pelvis	42,680	5%		Melanoma of the skin	36,120	4%
Non-Hodgkin lymphoma	41,730	5%		Non-Hodgkin lymphoma	32,950	4%
Oral cavity & pharynx	37,160	4%		Pancreas	26,240	3%
Leukemia	35,030	4%		Leukemia	25,270	3%
Liver & intrahepatic bile duct	30,610	4%		Kidney & renal pelvis	22,660	3%
All Sites	**856,370**	**100%**		**All Sites**	**878,980**	**100%**

Estimated Deaths

	Males			Females	
Lung & bronchus	83,550	26%		Lung & bronchus	70,500	25%
Prostate	29,430	9%		Breast	40,920	14%
Colon & rectum	27,390	8%		Colon & rectum	23,240	8%
Pancreas	23,020	7%		Pancreas	21,310	7%
Liver & intrahepatic bile duct	20,540	6%		Ovary	14,070	5%
Leukemia	14,270	4%		Uterine corpus	11,350	4%
Esophagus	12,850	4%		Leukemia	10,100	4%
Urinary bladder	12,520	4%		Liver & intrahepatic bile duct	9,660	3%
Non-Hodgkin lymphoma	11,510	4%		Non-Hodgkin lymphoma	8,400	3%
Kidney & renal pelvis	10,010	3%		Brain & other nervous system	7,340	3%
All Sites	**323,630**	**100%**		**All Sites**	**286,010**	**100%**

FIGURE 1. Ten Leading Cancer Types for the Estimated New Cancer Cases and Deaths by Sex, United States, 2018.
Estimates are rounded to the nearest 10 and cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder. Ranking is based on modeled projections and may differ from the most recent observed data.

year due to the time required for data collection, compila- applied to all 15 years of modeled data to estimate counts

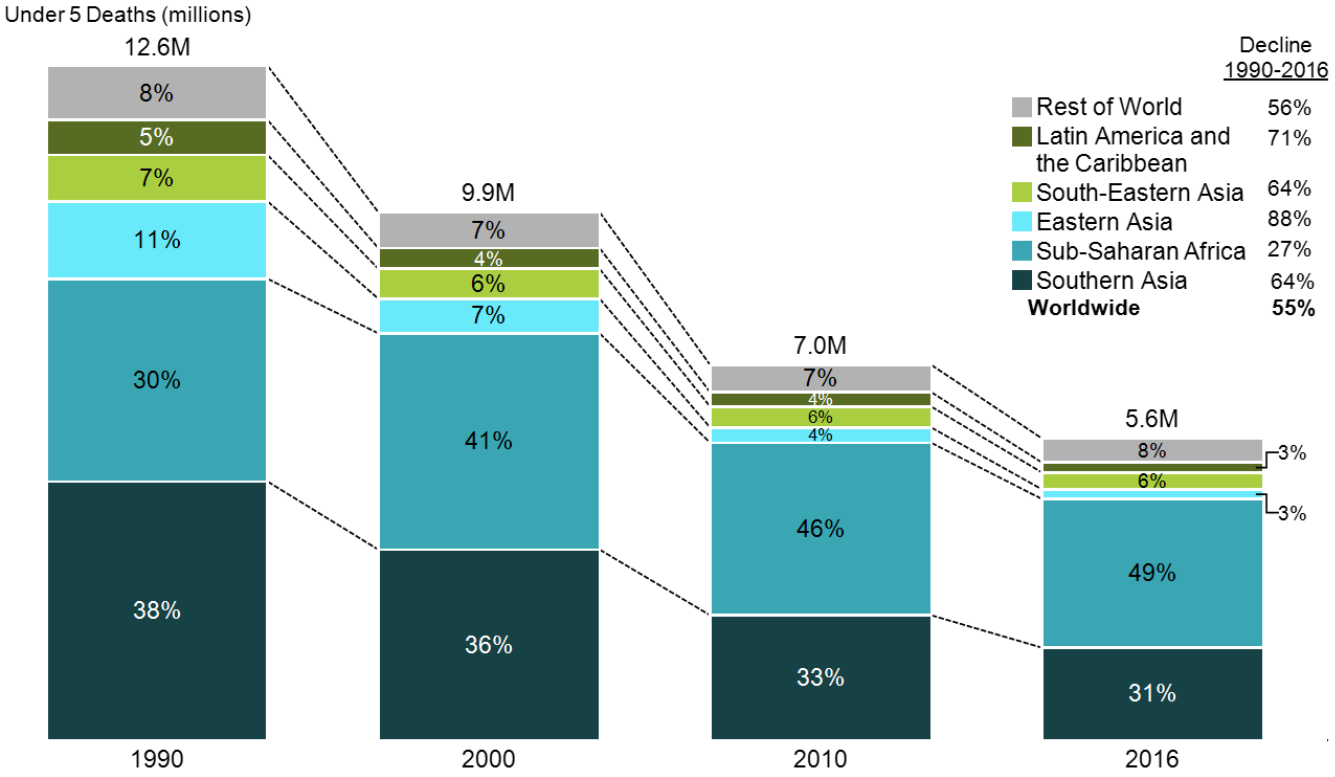
citation-264632605.txt | 1471-2466-14-134.pdf | 1055-9965.EPI-16-....pdf | 11.24.2018_PROS....docx | 12 Data Presentati....ppt | researchlecture.ppt | Show all

8:34 PM 10/23/2018

Stacked Bar Chart

Worldwide Decline in Child Mortality

Deaths in children under 5 have declined 55% worldwide from 1990-2016. In 2016, 80% of the 5.6M deaths occurred in Sub-Saharan Africa and Southern Asia.

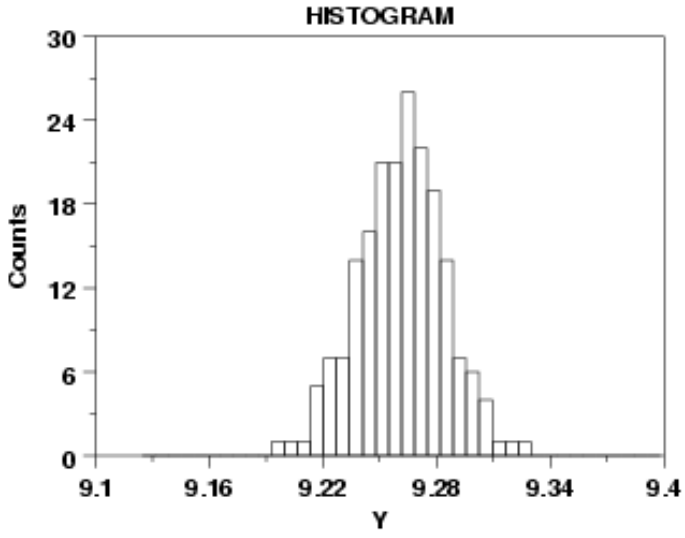


Numerical variables

- Variables that take numerical values:
  - Continuous when the variables can take any value within a certain range, i.e. include decimal values
    - E.g. weight (lbs), height (inches), blood pressure, age as measured on a continuous scale
  - Discrete: observations take only a certain numerical value
    - E.g. Number of prescriptions, number of pills per day or age in years

Measurement of central tendency

- Assumptions made around the distribution (normal or non-normal) of the data, determines which measure of central tendency to use
- Histograms are very powerful in summarizing the distribution of a univariate dataset
  - center (i.e., the location) of the data
  - spread (i.e., the scale) of the data
  - skewness of the data
  - presence of outliers
  - presence of multiple modes in the data.



Mean (average)

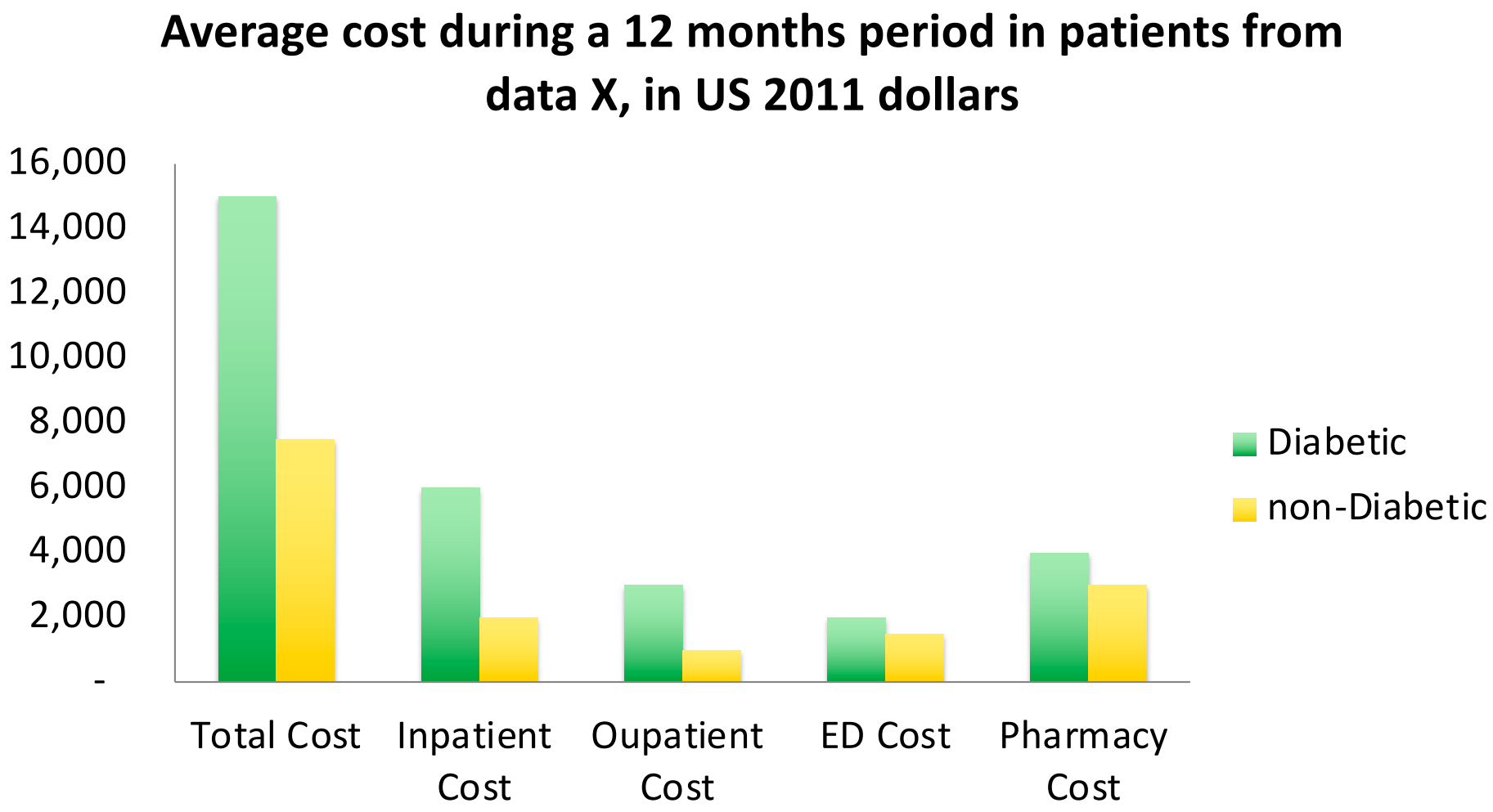
- Mean (average) is used when the data are assumed to be normally distributed

$$\text{Mean} = \frac{\sum x}{n}$$

  - Very sensitive to outliers and is not a true representation of the data if such outliers exist
  - e.g. the mean of this dataset {1,2,3,4,5,6,7,8} is 4.5
  - whereas the mean of {1,2,3,4,5,6,7,25} is 6.6.

"Add the numbers, divide by how many numbers you've added and there you have it-the average amount of minutes you sleep in class each day."

Bar Chart


Average cost during a 12 months period in patients from data X, in US 2011 dollars
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
-
Total Cost
Inpatient Cost
Oupatient Cost
ED Cost
Pharmacy Cost
Diabetic
non-Diabetic

Median & Range

- Median Describes the center of the data . By definition, 50% of the data resides below and the other 50% above the median
  - It is not impacted by outliers
- It mostly used to describe non-normally distributed data such as costs
  - e.g. the median of this dataset {1,2,3,4,5,6,7,8} is 4.5
  - And remains the same after replacing 8 with an outlier {1,2,3,4,5,6,7,25}
- Range is the difference in maximum and minimum values in a dataset

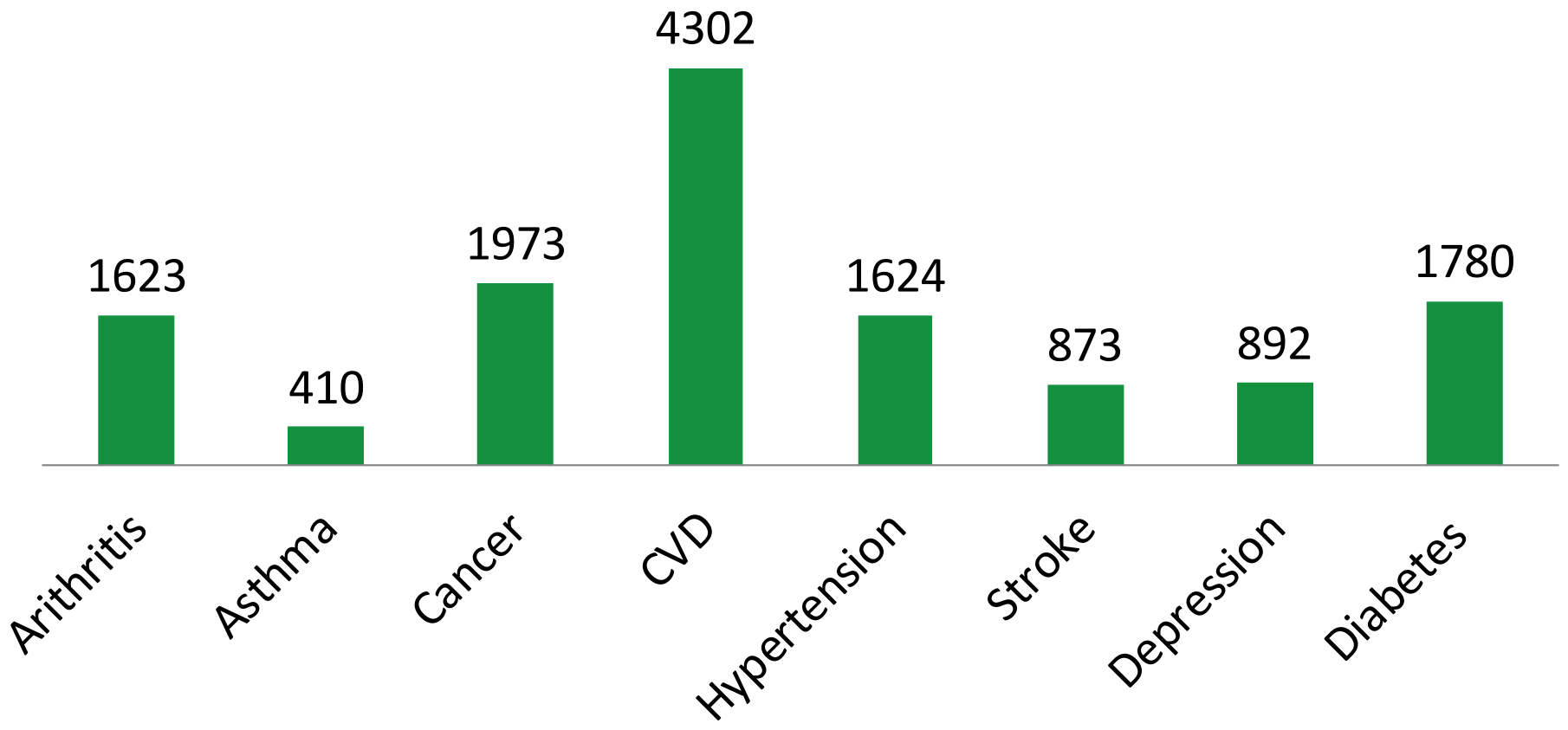
Tabular presentation

Chronic Condition	Medical Costs	Absenteeism Costs
	Median (Range)	
Arthritis	1,623 (206–11,490)	217 (28–1,791)
Asthma	410 (55–3,182)	48 (6–383)
Cancer	1,973 (227–13,614)	116 (15–916)
Cardiovascular disease, total	4,302 (411–26,062)	201 (23–1,396)
Diseases of the heart	2,376 (211–14,313)	76 (8–487)
Congestive heart failure	256 (18–1,702)	5 (0–42)
Coronary heart disease	1,340 (18–7,988)	58 (6–383)
Other heart diseases	765 (68–4,623)	13 (1–79)
Hypertension	1,624 (167–10,032)	78 (9–546)
Stroke	873 (86–5,959)	64 (7–451)
Depression	892 (120–6,728)	97 (13–861)
Diabetes	1,780 (212–12,095)	72 (9–544)

Trogdon JG, Murphy LB, Khavjou OA, Li R, Maylahn CM, Tangka FK, et al. Costs of Chronic Diseases at the State Level: The Chronic Disease Cost Calculator. Prev Chronic Dis 2015;12:150131. DOI: http://dx.doi.org/10.5888/pcd12.150131

Or presented graphically

Median State Specific Medical Cost in Millions, 2010 Dollars

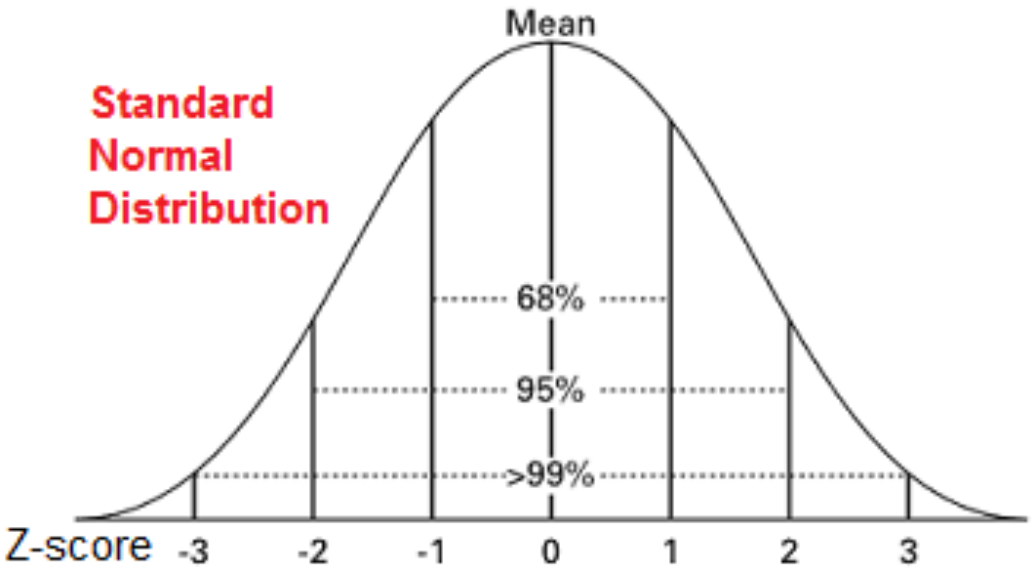


Standard deviation

- Standard deviation is most commonly used and defined as the average distance from the mean

$$SD=\sqrt{\frac{(x-\mu)^2}{n-1}}$$

- In the normal distribution 99% of the values should fall within 3 SDs, 95% within 2 SDs and 68% within 1 SD.



Standard error (SE) of the mean

- It is basically the standard deviation of all sample means if we were to repeatedly sample from the population
- SE of the mean describes how closely the sample mean approximates the population mean

$$SE=SD/\sqrt{n}$$

- SE inversely associate with the sample size.
- SE is useful when statistically comparing two means from two different samples to make inferences whether they represent similar populations

Confidence interval (CI)

- Gives a range and specifies probability that the true parameter value falls within this range
- Interpretation: If we were to repeat the experiment 100 times and compute 95% confidence intervals for each sample, then approximately 95 of the 100 CIs will include the true population estimate
- *In practice, we only sample once and make inferences about the true population mean, therefore, we may or may not capture the true mean*
- *95% probability that the confidence interval will contain the true population mean*

Means and CIs

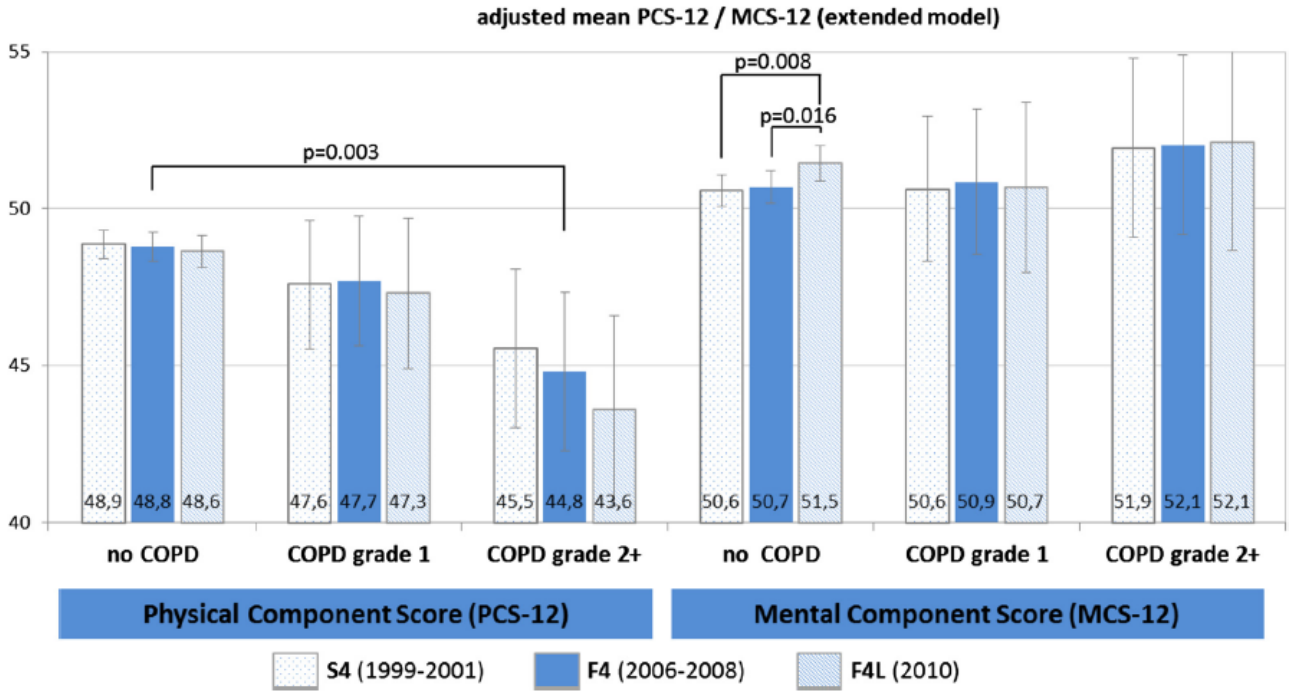


Figure 2 Adjusted mean PCS-12 and MCS-12 scores at S4, F4, F4L.

Wacker, Margarethe & Hunger, Matthias & Karrasch, Stefan & Heinrich, Joachim & Peters, Annette & Schulz, Holger & Holle, Rolf. (2014). Health-related quality of life and chronic obstructive pulmonary disease in early stages – longitudinal results from the population-based KORA cohort in a working age population. BMC pulmonary medicine. 14. 134. 10.1186/1471-2466-14-134.

Types of Analyses

- Univariate analyses
  - when describing one variable such as histograms or graphs.
- multivariate analyses
  - when more than one measurement is made on each observation
- Bivariate analyses
  - When only two measurements are made on a single observation

TTEST

- One sample TTEST is used to test whether the sample mean is different from a hypothesized value
  - The mean BMI of a diabetic sample of patients differs from 27.0

- Two **independent** samples TTEST is used to compare the means of a variable of interest between two groups
  - Test whether the mean BMI for diabetic female patients is similar to the mean BMI for diabetic male patients

normality assumption of the variable of interest should always be met before running the TTEST

Example (Independent TTEST)

	Surgical group ($n=22$)	Nonsurgical group ($n=20$)	P-value
Age	16.5 ± 2.1	14.8 ± 2.1	[a]**0.01**
Gender (male/female)	13/9	8/12	0.22
Interscan interval (days)	15.8 ± 17.9	11.2 ± 11.0	0.33
Time of imaging to surgery/endoscopy (days)	21.9 ± 18.7	19.5 ± 18.2	0.25
Time between diagnosis and surgery/endoscopy (months)	55.4 ± 72.9	15.8 ± 39.6	[a]**0.04**

Bold indicates significant P-value for a level of 0.05

[a] Calculated using two-sided t-test

Not-parametric tests

- Wilcoxon-Mann-Whitney test is used when the normality assumption is not met
  - The variable is at least ordinal
  - Widely used in skewed data for example health care resource utilization and cost data

Tests for categorical data

- Chi-square is used to test whether there is an association between categorical variables
  - E.g. prevalence of a certain disease by gender
  - Generates a contingency table with the assumption that each cell should include 5 or more observations
- Fishers exact test is used to test whether there is an association between categorical variables but relaxes the 5 observations per cell

Example

Table 3 Differences between surgical and nonsurgical groups on bowel US and MR enterography (MRE)

	Bowel US ($n=42$)			MRE ($n=42$)		
	Surgical group ($n=22$)	Nonsurgical group ($n=20$)	*P*-value	Surgical group ($n=22$)	Non-surgical group ($n=20$)	*P*-value
Bowel wall thickness (mm)	6.1 ± 1.8	4.7 ± 1.7	[a]***P* = 0.01**	9.1 ± 2.2	7.2 ± 2.8	[a]***P* = 0.02**
T2 ratio	N/A	N/A		4.6 ± 1.9	3.6 ± 1.1	[a]***P* = 0.03**
Loss of mural stratification	19/22 (86%)	10/20 (50%)	[b]***P* = 0.02** (OR = 6.3 [1.4-28.4])	N/A	N/A	
Vascularity score (0-3)	1.7	1.4	*P* = 0.3	N/A	N/A	
Fibrofatty proliferation (0-3)	2.2	1.6	[c]***P* = 0.04**	2.0	1.4	*P* = 0.06
Mesenteric edema (0-3)	N/A	N/A		2.2	1.2	[c]***P* = 0.001**
Stricture	4/22 (18%)	2/20 (10%)	*P* = 0.7 (OR = 2.0 [0.3-12.3])	13/22 (59%)	3/20 (15%)	[b]***P* = 0.005** (OR = 8.2 [1.8-36.4])
Abscess	5/22 (23%)	3/20 (15%)	*P* = 0.7 (OR = 1.6 [0.6-8.1])	8/22 (36%)	2/20 (10%)	*P* = 0.07 (OR = 5.1 [0.9-28.1])
Fistula	3/22 (14%)	2/20 (10%)	*P* = 1.0 (OR = 1.4 [0.2-9.5])	10/22 (45%)	4/20 (20%)	*P* = 0.1 (OR = 3.3 [0.8-13.3])

Values provided are means or counts with percentages. Odds ratios (OR) and 95% confidence intervals provided in brackets where appropriate. Bold indicates significant *P*-value for a level of 0.05

[a] Calculated using two-sided t-test

[b] Calculated using chi-square or two-sided Fisher exact test

[c] Calculated using Mann–Whitney *U* test

N/A not applicable

Take Away

- Data should be presented in an organized manner
- Data figures should be easy to understand
- Label axis and indicate source of the data
- Include footnotes as need for additional explanation
- Statistical tests have assumptions and depend on the type of data