V11 Multi-variate analysis

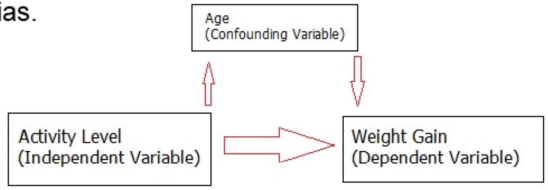
Program for today:

- What is a confounding variable?
- Ways how to avoid confounding effects.
- *Staphylococcus aureus* Africa project – analysis for confounding variables
- Which slides are relevant for final oral exam?
- Example questions.

In today‘s lecture, we will discuss the meaning of the term „confounding effect“, how these can be detected, and how these can be avoided by choosing an appropriate study design.

Confounding variables

A **confounding variable** is an "extra" variable that you didn't account for.
They can ruin an experiment and give you useless results.
They can suggest there is correlation when in fact there isn't.
They can even introduce bias.



In an experiment, the independent variable typically has an effect on your dependent variable.
E.g. if you are researching whether lack of exercise leads to weight gain, then lack of exercise is your independent variable and weight gain is your dependent variable.

Confounding variables are any other (extra independent) variable that also has a hidden effect on your dependent variable.

A confounding variable is closely related to both the independent and dependent variables in a study.

An independent variable represents the suppose cause, while the dependent variable is the supposed effect.

A confounding variable is a third variable that influences both the independent and dependent variables.

Confounding variables

Confounding variables can introduce **biases**.

Let's say you test 200 volunteers (100 men and 100 women) and you find that lack of exercise leads to weight gain.

One problem with this experiment is that it lacks any control variables, e.g. the use of placebos, or random assignment to groups.

One confounding variable could be **how much people eat**.

It is also possible that men eat more than women; this could also make **sex** a confounding variable.

Another possibility is **age**. E.g. if all of the women in the study were middle-aged, and all of the men were aged 16, **age** may have a direct effect on weight gain. That would make age a confounding variable.

This is just an example …

Case-control studies

There are various study designs that try to actively **exclude** or **control confounding variables**:

Case-control studies assign confounders equally to both groups, cases and controls.

E.g. if somebody wanted to study the cause of myocardial infarct and thinks that the age is a probable confounding variable, each 67-year-old infarct patient will be matched with a healthy 67-year-old "control" person.

In case-control studies, matched variables most often are the age and sex.

Drawback: Case-control studies are feasible only when it is easy to find controls.

https://en.wikipedia.org/wiki/Confounding

Suppose a case-control study attempts to find the cause of a given disease in a person who is 1) 45 years old, 2) African-American, 3) from Alaska, 4) an avid football player, 5) vegetarian, and 6) working in education. A theoretically perfect control would be a person who, in addition to not having the disease being investigated, matches all these six characteristics and also has no other diseases that the patient also does not have. Finding such a control would be an enormous task.

Cohort studies

A degree of matching is often realized by only admitting certain age groups or a certain sex into the study population.
This creates a **cohort** of people who share similar characteristics and thus all cohorts are comparable in regard to the possible confounding variable.

E.g. if age and sex are thought to be confounders, only 40 to 50 years old males would be involved in a cohort study that would assess the myocardial infarct risk in cohorts that either are physically active or inactive.

Drawback: In cohort studies, the over-exclusion of input data may lead researchers to define too narrowly the set of similarly situated persons for whom they claim the study to be useful. Then, other persons to whom the causal relationship does in fact apply may lose the opportunity to benefit from the study's recommendations.

Similarly, "over-stratification" of input data within a study may reduce the sample size in a given stratum to the point where generalizations drawn by observing the members of that stratum alone are not statistically significant.

https://en.wikipedia.org/wiki/Confounding

Wikipedia explains: In statistics, stratified sampling is a method of sampling from a population which can be partitioned into subpopulations.

Double blinding

Double blinding conceals (hides) from the trial population and the observers the experiment group membership of the participants.

By preventing the participants from knowing if they are receiving treatment or not, the placebo effect should be the same for the control and treatment groups.

By preventing the observers from knowing of their membership, there should be no bias from researchers treating the groups differently or from interpreting the outcomes differently.

https://en.wikipedia.org/wiki/Confounding

In a **double-blind** study (dt. Blindstudie), participants and experimenters do not know who is receiving a particular treatment.

"Double" specifies that both the participants AND the staff conducting the study (e.g. medical doctors) do not know to which group the participants belong.

Randomized controlled trial

This is a method where the **study population is divided randomly** in order to mitigate the chances of self-selection by participants or bias by the study designers.
Before the experiment begins, the testers will assign the members of the participant pool to their groups (control or intervention) using a randomization process such as the use of a random number generator.

E.g. in a study on the effects of exercise, the conclusions would be less valid if participants were given a choice if they wanted to belong to the control group which would not exercise or the intervention group which would be willing to take part in an exercise program.
The study would then capture other variables besides exercise, such as pre-experiment health levels and motivation to adopt healthy activities.

From the observer's side, the experimenter may choose candidates who are more likely to show the results the study wants to see or may interpret subjective results (more energetic, positive attitude) in a way favorable to their desires.

https://www.medicinenet.com/script/main/art.asp?articlekey=39532:

Randomized controlled trial: (RCT) A study in which people are allocated at random (by chance alone) to receive one of several clinical interventions. One of these interventions is the standard of comparison or control. The control may be a standard practice, a placebo ("sugar pill"), or no intervention at all. Someone who takes part in a randomized controlled trial (RCT) is called a participant or subject. RCTs seek to measure and compare the outcomes after the participants receive the interventions. Because the outcomes are measured, RCTs are quantitative studies.

In sum, RCTs are quantitative, comparative, controlled experiments in which investigators study two or more interventions in a series of individuals who receive them in random order. The RCT is one of the simplest and most powerful tools in clinical research.

Stratification

As in the example just mentioned, physical activity is thought to be a behavior that protects from myocardial infarct; and age is assumed to be a possible confounder.

The data sampled is then **stratified by age group** – this means that the association between activity and infarct would be analyzed per each age group.

If the different age groups (or age strata) yield much different risk ratios, age must be viewed as a confounding variable.

There exist statistical tools that account for stratification of data sets.

https://en.wikipedia.org/wiki/Confounding

https://asq.org/quality-resources/stratification

Stratification is defined as the act of sorting data, people, and objects into distinct groups or layers. It is a technique used in combination with other data analysis tools. When data from a variety of sources or categories have been lumped together, the meaning of the data can be difficult to see. This data collection and analysis technique separates the data so that patterns can be seen.

Here are examples of different sources that might require data to be stratified:

Equipment, Shifts, Departments, Materials, Suppliers, Day of the week, Time of day, Products

STRATIFICATION PROCEDURE

- Before collecting data, consider which information about the sources of the data might have an effect on the results. Set up the data collection so that you collect that information as well.

- When plotting or graphing the collected data on a scatter diagram, control chart, histogram, or other analysis tool, use different marks or colors to distinguish data from various sources. Data that are distinguished in this way are said to be "stratified."

- Analyze the subsets of stratified data separately.

Review (V1): *S. aureus* in Germany vs. Africa: StaphNet

6 study sites each collected 100 isolates of healthy volunteers and 100 of blood culture or clinical infection sites.

Aim

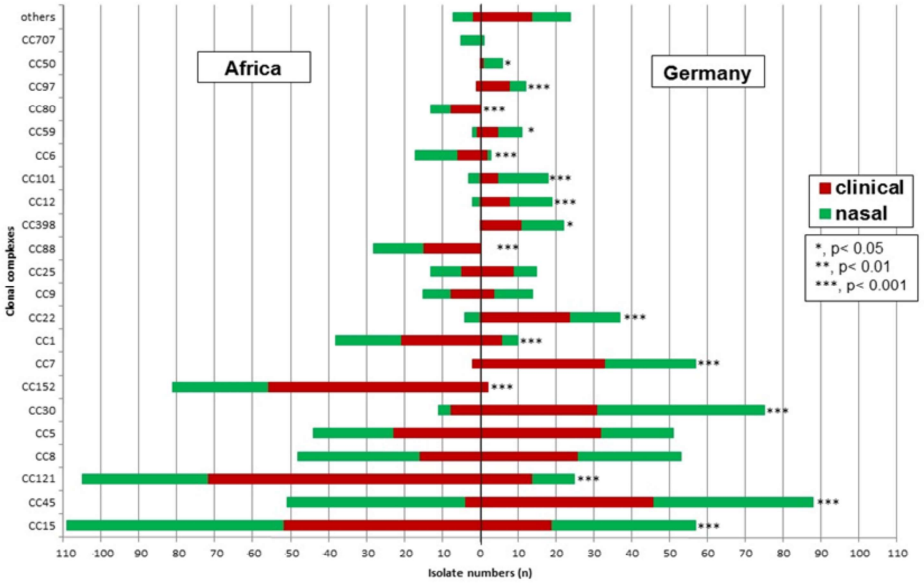
microbiological and molecular characterization of African *S. aureus* isolates

by DNA microarray analysis including clonal complex analysis

supplemented by Whole Genome Sequencing

Now, we will look back at the example that we discusses in the first 2 lectures, the collection of bacterial Staphylococcus aureus samples in three African countries and in three German university hospitals.

Review (V1): Distribution of clonal complexes

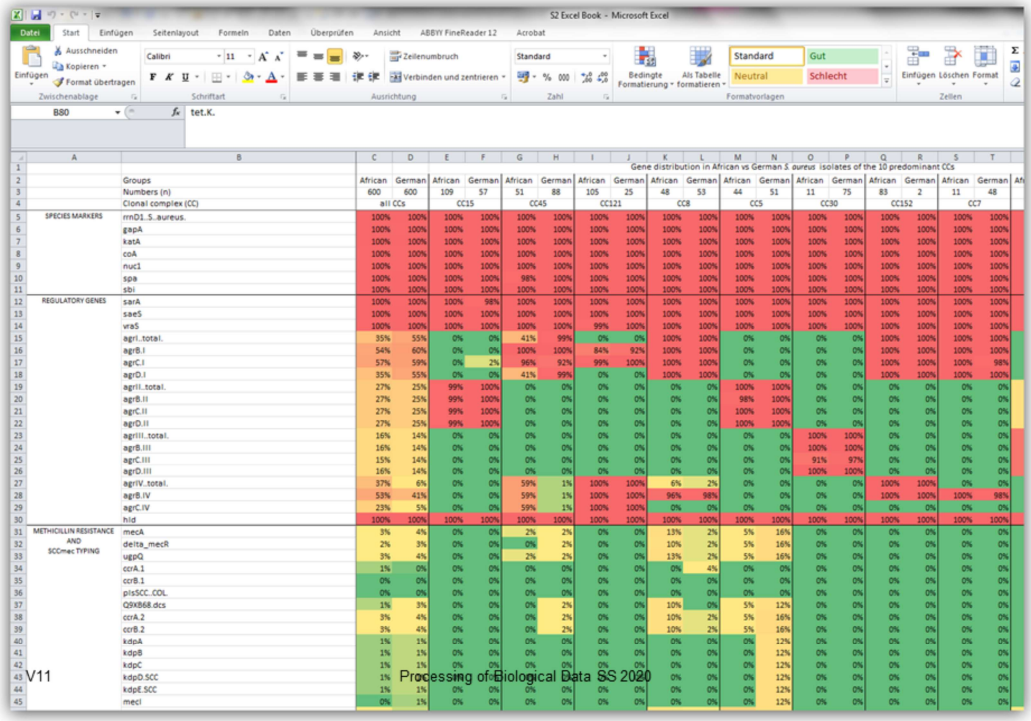


Some clonal complexes more prevalent in Africa, others predominant in Germany.

V11 Processing of Biological Data SS 2020 10

This is an overview to which clonal complexes the samples belong.

Review (V1): Activitity of individual probes for CCs



This slide summarizes the results of the hybridization against the DNA microarray.

The columns denote what fraction of the African or German samples belonging to a particular clonal complex (third row) contain certain genes.

Red: practically all samples hybridize against this probe.

Dark Green: 0% show positive hybridization,

Light green : 1-2% show positive hybridization,

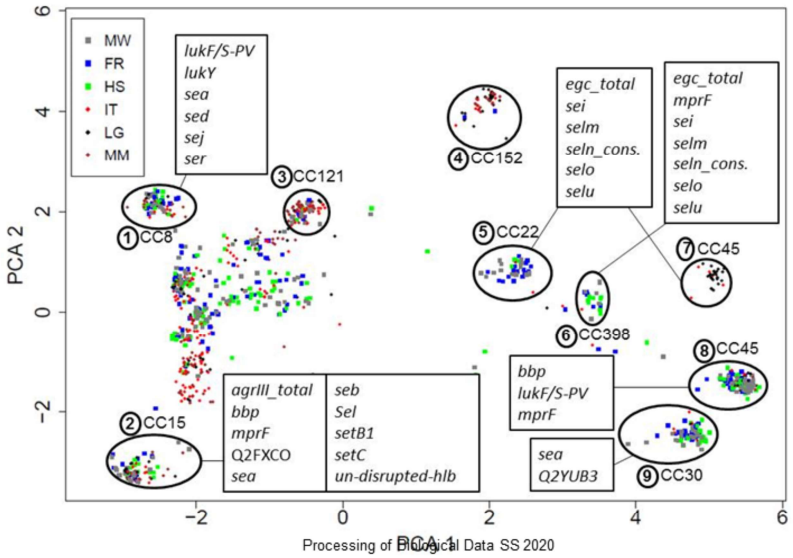
Yellow and orange: increasing fractions show positive hybridization.

Review (V1): Principal component analysis of 1200 strains

Input data: binary matrix of MA data; dimension 1200 x 334 probes

PCA identifies local clusters that are characteristic

for particular clonal complexes



This slide shows a principal component analysis of the microarray hybridization results (data in table on previous slide).

Circled clusters often contain only members of one clonal complex.

Staphylococcus aureus data from Africa project (V1)

Site	# of cases below 1 year	# of cases 1 to 5 years	# of cases 6 - 25 years	# of cases 26 – 65 years	# of cases above 66 years
Africa + Germany (clinical)	**88**	**109**	**90**	**225**	**88**
Africa + Germany (commensal)	19	34	363	175	9
Africa (clinical)	**86**	**106**	**53**	**54**	**1**
Africa (commensal)	17	**34**	156	89	**4**
Germany (clinical)	**2**	**3**	**37**	**171**	**87**
Germany (commensal)	2	**0**	207	86	**5**

Age distribution was **heavily skewed**:

many small kids / babies in Africa – few seniors in Africa

very few small kids / babies in Germany – many seniors in Germany

Did this affect the analysis + interpretation?

V11 Processing of Biological Data SS 2020 13

This is something I did not mention in the lectures #1 and #2: the age distribution between African and German samples is quite different.

We were worried whether this would affect the outcomes of our study that focused on the analysis of bacterial samples, not on the human carriers.

Are elderly people preferably colonized by different bacterial strains than young people?

In principle, one could expect that the total numbers of samples in Africa vs. Germany would be biased by the higher life expectation in Germany than in Africa.

However, the fraction of commensal samples from elderly people is quite low in both continents (1 vs. 5).

In contrast, the ratios are very different for the clinical samples.

Africa had many clinical cases for infants and small children reflecting the problematic health situation in Africa.

Many African clinical cases were apparently due to traffic accidents, after which the victims had to be taken to far away hospitals, which could take days.

Germany had many clinical cases for elderly people reflecting their higher susceptibility toward infections.

Analyze whether age is a confounding variable

To test whether age is a **confounding variable**, one can compare the results from simple linear regression with those from multiple linear regression.

The principal difference between these two types of regression models is the number of explanatory variables:

(1) the simple linear regression (SLR) model uses only one dependent variable ***y*** and one explanatory variable ***x***: $y = a + b \cdot x$

In our case, ***y*** stands for the binary output from the Alere-chip experiment for a particular gene. ***y*** therefore has values of 0 or 1.

With the binary variable ***x*** we could encode the sites Africa ($x = 0$) / Germany ($x = 1$). ***a*** and ***b*** are weights estimated by the model.

Generally SLR tries to find such weights (values for ***a*** and ***b***) so that the difference between the estimated ***y*** and actual ***y*** will be the smallest.

We will test whether age is a confounding variable by performing different types of linear regression analysis.

As a reminder, linear regression yields an optimal fit of a line y = a + b . x to the data.

Analyze whether age is a confounding variable

(2) the multiple linear regression model also has one dependent variable ***y*** but more than one explanatory variable

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n$$

As above, ***y*** will be the Alere-chip entry for a gene with value 0 or 1.

The **site**, **clinical/commensal** and **age** categories will be used as explanatory variables .

We will compare the results of a linear fit to a multiple linear regression model against several variables.

Steps of testing age categories for confounding

(1) Estimate a linear regression model for the dependent variable and one or more explanatory variables.

(2) Repeat step 1 with age categories added as further explanatory variable.

(3) Compare the weights obtained in steps 1 and 2.

As a rule of thumb, if the weight (s) (regression coefficient(s)) from step 1 changes by more than 10%, then the additional variable (here: age) may be considered as a **confounder**.

By following these steps, one can test for every significant finding (e.g. gene association) whether age is a confounder.

Reasons for this could be e.g. a significant imbalance in the distribution of age among samples.

Obviously, the two different fit approaches will not give „the same“ result.

What degree of difference should be considered as an „alarm“ sign of a possible confounding effect?

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6447501/

states „A statistical approach to covariate selection … is what is sometimes called the “change-in-estimate” approach. In this approach covariate selection decisions are made based upon whether inclusion of a covariate changes the estimate of the causal effect for the exposure by more than some threshold, often 10%.”

Case study

Case study: test whether age categories are a confounding variable for the 2 genes lukS.PV and sdrC..total

Reason for selecting these genes:
these 2 genes have very different frequencies in African vs German sites as well as in clinical vs commensal samples.

Africa was encoded as $x_1 = 0$ and Germany as $x_1 = 1$.

Clinical samples were encoded as $x_2 = 1$ and commensal with $x_2 = 0$.

Age categories were encoded from $x_3 = 1$ to 5.

We performed this test for two genes lukS.PV and sdrC..total that showed very imbalanced frequencies.

Our reasoning was that if we could not find an age effect here, then we could argue that the age imbalance did not significantly affect the results we reported.

Multiple linear regression model for the lukS.PV gene

The Alere result for this gene for different samples is the dependent variable.
The site affiliation + clin/com values are explanatory variables.

The table lists the dependent (lukS.PV) and explanatory (Africa_value, clin_com_value) variables for 10 samples out of 1200 samples.

#	samples	lukS.PV	Africa_value	clin_com_value
1	FR-B001	0	0	1
2	FR-B003	0	0	1
3	FR-B004	0	0	1
4	FR-B005	0	0	1
5	FR-B007	0	0	1
6	FR-B008	0	0	1
7	FR-B009	0	0	1
8	FR-B010	0	0	1
9	FR-B011	0	0	1
10	FR-B012	0	0	1

Since all these samples are from a German site, the Africa_value = 0.
Also, all samples are clinical (clin_com_value = 1).

The table shows a small piece of the used data. In total, it contains 1200 rows.

Our first regression model relates the value of the lukS.PV column (0 or 1) to the values of the Africa_value column (0 or 1) and the clin_com_value (0 or 1).

IukS.PV

Application of linear regression determines optimal weights w_1, w_2, w_3 .

For every sample we get IukS.PV = $w_1 + w_2 \cdot$ Africa.value + $w_3 \cdot$ clin com value .

For the first sample FR-B001, the formula would be $0 = w_1 + w_2 \cdot 0 + w_3 \cdot 1$.

Results from multiple linear regression (coefficients marked in bold):

	Estimate	Std. Error	t value	Pr(>|t|)
w_1 for intercept	-0.07250	0.01781	-4.070	5e-05 ***
w_2 for Africa_value	**0.42833**	0.02057	20.825	<2e-16 ***
w_3 for clin_com_value	**0.19500**	0.02057	9.481	<2e-16 ***

In other words, the following model is estimated:
IukS.PV = -0.07 + 0.42833 · Africa_value + 0.195 · clin_com_value

t value : equal to coefficient (estimate) divided by the standard error.
Pr(>|t|) : p-value = probability of seeing a result as extreme in random data.

The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line.

Although the model is very simplistic, the std. error is quite small.

IukS.PV

We then added a further variable "age category" with weight w_4 to the model.

IukS.PV = w_1 + w_2 · Africa.value + w_3 · clin com value + w_4 · age

	Estimate	Std. Error	t	value Pr(>|t|)
(Intercept)	0.06211	0.04559	1.362	0.17333
Africa_value	**0.39077**	0.02360	16.556	< 2e-16 ***
clin_com_value	**0.19470**	0.02049	9.503	< 2e-16 ***
age	**-0.03618**	0.01129	-3.206	0.00138 **

IukS.PV =
0.06211 + 0.39077 · Africa value + 0.19470 · clin com value - 0.03618 · age

V11 Processing of Biological Data SS 2020 20

This is the same fit when we added age as a third variable.

Actually, we did not use the actual age of the cases, but grouped the cases into the 5 age categories listed on slide 13: below 1 year, 1 to 5 years, 6 - 25 years, 26 – 65 years, above 66 years

The categories were encoded as 0 to 4 or 1 to 5.

lukS.PV

This result shows
(a) that the age category has a very small impact (its own weight is close to 0) and
(b) the two other weights (for the site and clin/com) did not change much.

E.g. the weight of the Africa_values changed in relative terms by :

$$\frac{(0.42833 - 0.39077)}{0.42833} \cdot 100\% = 8.8\%$$

The weight of clin_com_value changed by only **0.15%.**

Both values are smaller than 10% (rule of thumb).

Conclusion:
There is no (clear) statistical evidence that age acts as a confounding variable.

The addition of an age variable had a very small effect on the linear regression of lukS.PV status.

Same analysis for gene sdrC_total

Before adding age categories:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.02083	0.0125	81.711	< 2e-16 ***
Africa_value	-0.12833	0.0144	-8.896	< 2e-16 ***
clin_com_value	-0.05833	0.0144	-4.044	5.6e-05 ***

After adding age categories:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.975445	0.0321	30.407	< 2e-16 ***
Africa_value	-0.115667	0.0166	-6.964	5.44e-12 ***
clin_com_value	-0.058232	0.0144	-4.039	5.71e-05 ***
age-category	0.012198	0.0079	1.536	0.125

Weight of Africa_value changed by **9.87%**, weight of clin_com_value changed by **0.17%**

Also for sdrC, the age category got a small weight and the probability (p-value) of a random-effect is actually quite high.

In this case, the weight of the Africa_value changed by almost 10%. We still considered this close enough to the simpler fit without age category.

Conclusion

There is no evidence from our preliminary analysis for the genes lukS.PV and sdrC..total that age acts as a confounder in the association of genes with invasiveness and site affiliation.

We wrote in our manuscript:
"The discrepancy in population age between the German and African cohort potentially biases the 'true' distribution of clones and genes between isolates from the different geographic regions …
[but] application of a multiple linear regression model for the detection rate of Panton-Valentine leucocidin genes failed to provide evidence that age acts as a confounding variable"

Ruffing et al. Sci. Rep. 7, 154 (2017)

We are smart ☺

In our manuscript, we first acknowledge openly that age may be confounding factor.

Then, we state that a multiple linear regression model did not provide evidence FOR it. But we also did not exclude that such an effect may exist.

The reviewers of our study were satisfied.

Diabetes/HIV as confounding variables

Next, we tested using Fisher's exact test whether
(a) diabetes and HIV have similar frequencies in the total groups of African and German samples and
(b) whether diabetes and HIV have similar frequencies in selected groups of African and German individuals carrying particular clonal complexes.

The Fisher test considers the distribution provided in a 2 x 2 table (enclosed red).

	Africa	Germany	Row Total
HIV+	a	b	a + b
HIV-	c	d	c + d
Column Total	a + c	b + d	a + b + c+ d = n

The probability of obtaining any such set of values is given by the hypergeometric distribution:

$$p - value = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

V11 Processing of Biological Data SS 2020 24

This is additional data and analysis that we did not report in our study.

We also tested whether the reported findings were affected in a possible imbalance in the occurrence of diabetes and HIV.

One convenient way to test this is Fisher‘s exact test.

Analysis of HIV co-infection

First, we will test the null hypothesis that "HIV is equally distributed in African and German samples".

(a) For all African samples and all German samples we obtain the following dependencies of HIV carriers (HIV+) and of individuals without approved HIV status (you may say non-carriers) (HIV-):

	Africa	Germany
HIV+	41	0
HIV-	315	586

The p-value obtained for this table can be interpreted as the sum of evidence provided by the observed data—or any more extreme table—for the null hypothesis that "there is no difference in the proportions of HIV carriers among the African and German individuals tested in our study".

The smaller the value of p, the greater the evidence for rejecting the null hypothesis.

Very sadly, Africa is strongly affected by HIV infections. This is also reflected in the study cohort.

In the German cohort, there was no HIV case included. This may either be co-incidence or result from better monitoring of the population and exclusion of HIV infected cases.

Analysis of HIV co-infection

For the data shown above,

$$p - value = \frac{(41+0)!(315+586)!(41+315)!(0+586)!}{41!0!315!586!942!} = 1.03838\text{e-}18$$

Thus, there is very strong evidence from the observed frequencies that African and German individuals **are not equally likely to be HIV carriers**.

The same Fisher‘s test was applied.

Indeed, the very small p-value is evidence that the HIV status differs significantly in African and German cases.

Analysis of diabetes co-infection

Similarly, we can obtain Fisher's exact p-value for the distribution of **diabetes** among African and German samples.

	Africa	Germany
diab+	4	68
diab-	475	526

p-value = 3.73425e-14

Also, here, the null hypothesis of a similar distribution is **strongly rejected** suggesting the prevalence of diabetes in individuals from Germany compared to individuals from Africa.

Of course, we can trace this imbalance back to the **difference in age categories** of the two groups.

TABLE 1

Prevalence (95%-confidence intervals) and incidence of type 2 diabetes for male and female policyholders of German statutory health insurance funds over 40 years of age (2009, 2010)*

Age group	Prevalence (%) 2009	Prevalence (%) 2010	Age	Incidence/1000 py R	Incidence/1000 py R–15 %	Incidence/1000 py R+15 %
Men						
40–49 years	1.5 [1.5; 1.6]	1.6 [1.6; 1.7]	45 years	4	4	4
50–59 years	5.5 [5.4; 5.5]	5.7 [5.6; 5.7]	55 years	9	9	9
60–69 years	14.0 [14.0; 14.1]	14.5 [14.4; 14.5]	65 years	18	18	19
70–79 years	21.1 [21.1; 21.2]	21.9 [21.8; 21.9]	75 years	24	23	26
80–89 years	25.1 [25.0; 25.2]	26.3 [26.2; 26.4]	85 years	29	25	32
90–99 years	23.4 [23.1; 23.6]	24.1 [23.9; 24.4]	95 years	26	17	33
≥ 100 years	17.4 [15.3; 19.6]	16.5 [14.4; 18.6]	105 years	17	3	31
Over 40 years	7.03 [7.02; 7.04]	7.41 [7.40; 7.42]	Over 40 years	16	15	17

https://www.aerzteblatt.de/int/archive/article/175344/

https://www.aerzteblatt.de/int/archive/article/175344/The-prevalence-and-incidence-of-diabetes-in-Germany-an-analysis-of-statutory-health-insurance-data-on-65-million-individuals-from-the-years-2009-and-2010

reports that the incidence of diabetes in Germany is roughly 10% of the population.

In Africa, the prevalence is estimated as about 4% of the population, see https://idf.org/our-network/regions-members/africa/welcome.html

So this difference would not explain the large imbalance observed in the confusion matrix.

The proper answer is provided by the prevalence in different age groups, see right table.

There is a steep rise from 1.6% for men between 40-49 years old to 26.3% for men between 80-89 years old.

Hence, the observed age imbalance strongly affects the prevelance of diabetes in African vs. German cases.

HIV/diabetes in individuals with selected CCs

Next, we tested the distribution of HIV/diabetes in individuals carrying *S. aureus* from selected clonal complexes (CC15, CC45, CC121, CC30 which showed significant imbalance in German/African samples).

These are the results (tables + p-values from Fisher's exact test)

RF_HIV

CC15	Africa	Germany
hiv+	4	0
hiv-	65	57
p-value	0.126	0.25 (after correction for false discovery rate (FDR))

CC45	Africa	Germany
hiv+	1	0
hiv-	40	87
p-value	0.320	0.42 (FDR-corrected)

As shown before, there exists strong and statistically significant imbalances of HIV and diabetes in the study cohort.

But does this also affect the findings reported by us?

Our focus was not placed on the individuals and their infection status, but on the bacterial strains colonizing or infecting them.

Hence, we repeated the same analysis shown before for the major subpopulations of clonal complexes.

Now it turns out that the imbalances were mostly statistically insignificant.

Because we performed multiple tests, we had to apply an FDR correction.

HIV/diabetes in individuals with selected CCs

CC121	Africa	Germany
hiv+	11	0
hiv-	40	24
p-value	0.0132	**0.05** (FDR-corrected)

CC30	Africa	Germany
hiv+	0	0
hiv-	7	75
p-value	1	1 (FDR-corrected)

Colored red are those scenarios with FDR-corrected p-value less or equal to the significance threshold of 0.05.

HIV/diabetes in individuals with selected CCs

RF_CCSI_Diab_mel

CC15	Africa	Germany
diab+	0	1
diab-	88	56
p-value	0.393	0.52 (FDR-corrected)

CC45	Africa	Germany
diab+	0	12
diab-	47	75
p-value	0.0081	**0.03** (FDR-corrected)

CC121	Africa	Germany
diab+	0	1
diab-	57	24
p-value	0.305	0.52 (FDR-corrected)

CC30	Africa	Germany
diab+	0	7
diab-	9	68
p-value	1	1 (FDR-corrected)

Interpretation

In most cases, there is no evidence based on our data to reject the null hypothesis of assuming a similar distribution of HIV and diabetes carriers among African and German samples belonging to **particular clonal complexes.**

The only exceptions to this are CC45 (diabetes – p=0.008/q=0.03)
and CC121 (HIV – borderline p=0.013/q=0.05).

Therefore, we summarized in an internal report of the project team:
- we observed statistically significant imbalances in the frequencies of all these clonal complexes XXX, YYY ... between African and Germany.
- we tested based on Fisher's exact test that these imbalances were not due to an imbalance of HIV and diabetes carriers in both groups.
- The only exceptions to this are CC45 (diabetes) and CC121 (HIV) where such associations cannot be ruled out, but are only weakly supported by the data.

We did not even mention the outcome of these checks in the manuscript because we believe that the reported results are not effected by the HIV/diabetes status of the individuals.

Summary

- It is important to think about the occurrence of possible confounding effects
- Multi-variate vs. single-variate analysis reveals possible confounding effects
- Choosing an appropriate study design reduces the likelihood that confounding effects may occur.

Relevant slides for oral exam on July 27-31, 2020

Lecture	Slides
1	14, 18, 19, 26-30, 35-40
2	4-7, 9, 13-15, 18, 22-23, 26-32
3	6-7, 13-16, 18-23, 32
4	11-17, 19-23, 42-43
5	4-6, 17, 19-22, 33-34
6	3 (only Hi-C), 7-11, 14-23, 26-37
7	16, 18-24, 35-39
8	4-5, 8
9	10-25, 39-43
10	5, 8-14, 25- 27, 33-35
11	2, 4-8, 13-31
Assignments	Ass.#2 (full), Ass. #3 (full)

I will not ask questions about the other slides & assignments.

Example questions

(1) What types of methods exist to perform integrated analysis of multiple omics datasets?
(*Answer: multi-staged approaches and meta-dimensional approaches*)

To which category does the tool iCluster belong?
(*Answer: it is a meta-dimensional method – uses model-based integration*)

How does it work?

(2) How does the watershed algorithm work?

(3) What is the connection between PCA and singular value decomposition?

Can one perform PCA without SVD?
(*Answer: yes, e.g. by geometric construction, see slide 29 in V1*).