2nd Symposium on Molecular Radiotherapy Dosimetry:

The future of theragnostics

November 13th - 15th 2025, Athens, Greece

Hosted by

SIG RADIONUCLIDE INTERNAL DOSIMETRY

1969 THE HELLENIC ASSOCIATION OF MEDICAL PHYSICISTS (HAMP)

EFOMP EUROPEAN FEDERATION OF ORGANISATIONS FOR MEDICAL PHYSICS

Statistics In MRT Dosimetry

Convenors: M Brambilla, M Cremonesi & L Strigari

Statistics in MRT Dosimetry

WHY? Statistical analysis is of utmost importance in most of the dosimetry studies for MRT.
However, many publications and clinical trials also show inappropriate application of statistical methods,
do not consider the hypothesis needed, do not report essential details, etc.

This may lead to **misinterpreting the meaning or robustness of the results**, sustaining erroneous conclusions.

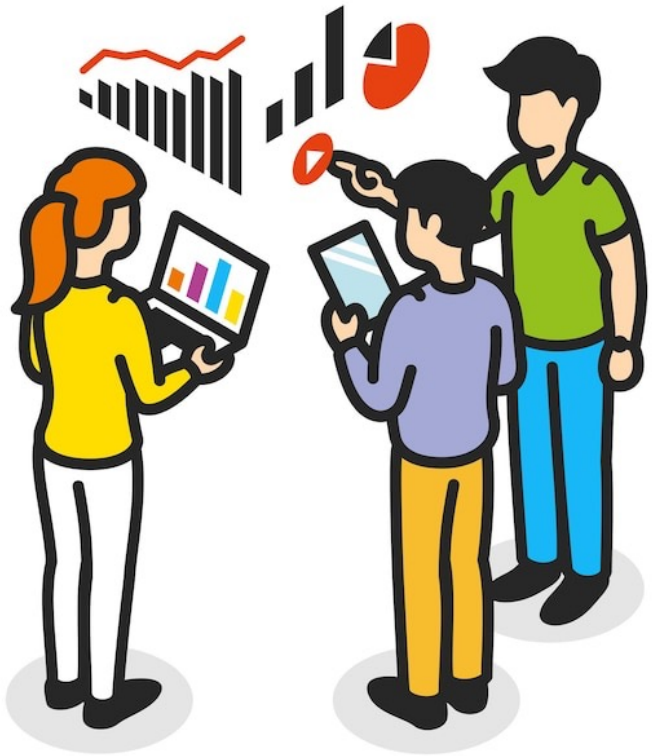
Typically, the statistical approach is based on what was applied to previous publications, without a critical view

Appropriate use of statistical methods: «a big deal» only in dosimetry for MRT?

Statistics in MRT Dosimetry

AIM

With the intention of **promoting a more critical view and awareness**, we will focus on typical issues in dosimetry, revising representative examples, summarising the theory of the statistics involved, making constructive observations, and, hopefully, **openly interacting** with the audience.



Appropriate use of statistical methods: «a big deal» only in dosimetry for MRT?

Outline

- Agreement

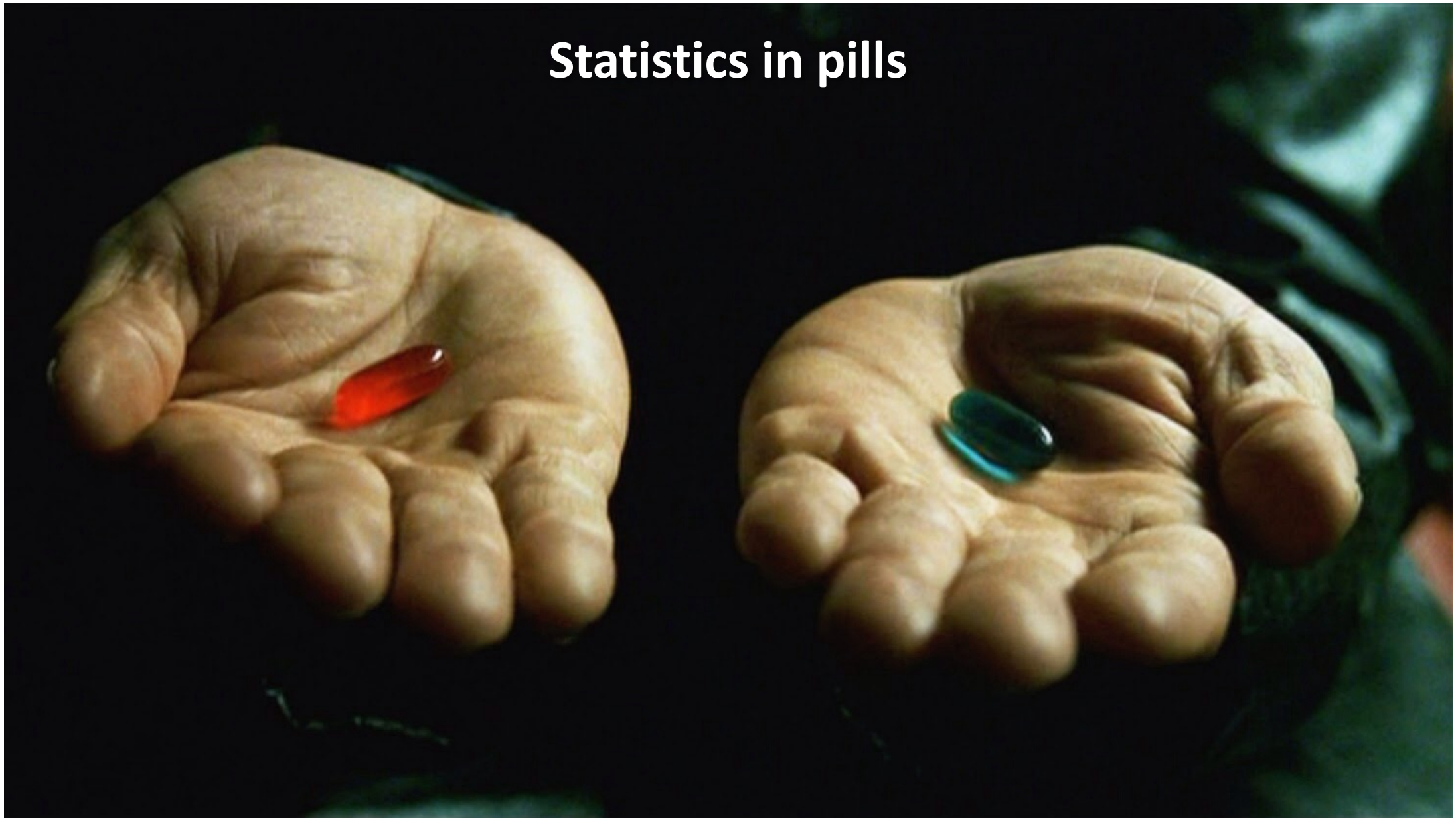
  Bias
  Limits of agreement
  Lin's Concordance coefficient

- Correlation, Regression (dose –response relationship)

  Explanation: R^2, standardized regression coefficients
  Prediction: Regression, unstandardized regression coefficients

- Building NTCP curves, Logistic Regression

- **OPEN DISCUSSION**


Statistics in pills

2nd Symposium on Molecular Radiotherapy Dosimetry:

The future of theragnostics

November 13th - 15th 2025, Athens, Greece

Hosted by

SIG RADIONUCLIDE INTERNAL DOSIMETRY

THE HELLENIC ASSOCIATION OF MEDICAL PHYSICISTS (HAMP)

EFOMP EUROPEAN FEDERATION OF ORGANISATIONS FOR MEDICAL PHYSICS

Statistics In MRT Dosimetry

Agreement, bias, correlation, regression, curve fitting and others

Marco Brambilla

Department of Medical Physics,

University Hospital “Maggiore della Carità

How fragile is medical research?

Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2(8):e124.
Using a Bayesian framework, he demonstrated that in fields with small sample sizes, low prior probabilities, and high flexibility in design and reporting, the majority of positive findings are likely false positives.

Prasad V, et al. A Decade of Reversal: An Analysis of 146 Contradicted Medical Practices. Mayo Clin Proc. 2013;88(8):790-798.
identified 146 cases of “medical reversals”—treatments once thought effective but later proven ineffective or harmful—within just ten years.

Grimes DR, et al. Towards replicability and sustainability in cancer research. Nat Cancer. 2024;5:609–616.
only 11% of landmark cancer biology experiments could be reproduced under rigorous testing.

Possamai A, et al. Inclusion of Retracted Studies in Systematic Reviews and Meta-analyses. JAMA Intern Med. 2025.
35% of meta-analyses changed their conclusions by at least 10% once retracted studies were removed.

Xu S, et al. Investigating the impact of trial retractions on evidence synthesis. BMJ. 2025;389:e082068.
showed how retracted trials contaminated guidelines, leading to flawed clinical recommendations

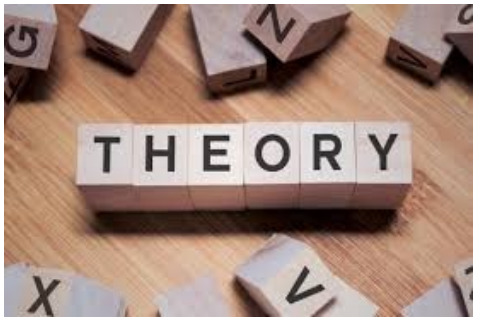
Cobey KD, et al. Biomedical researchers’ perspectives on reproducibility. J Clin Epidemiol. 2024;163:58–68.
A 2024 international survey of over 1,600 biomedical scientists reported that 72% believe there is a reproducibility crisis, and 62% blame pressure to publish as a key driver.

Agreement

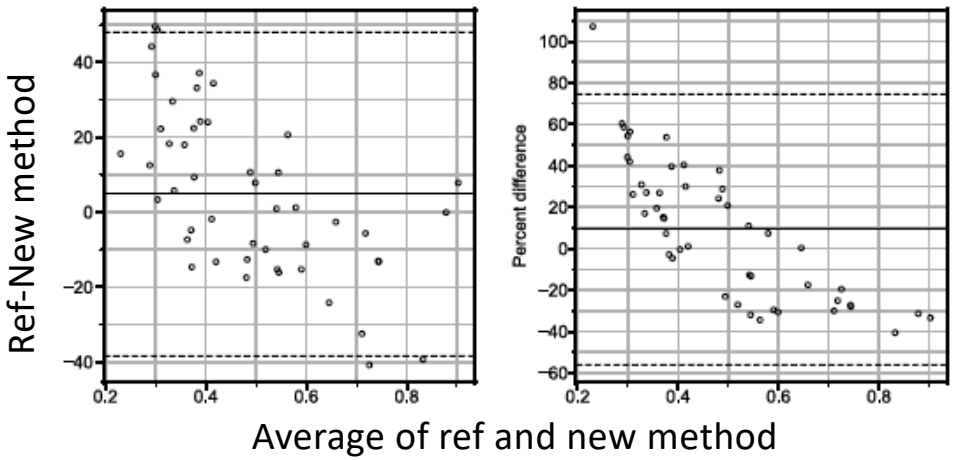
- Bias
- Limits of agreement
- Lin's Concordance coefficient

1. παραδείγματα

2. θεωρία



Case 1 – Agreement

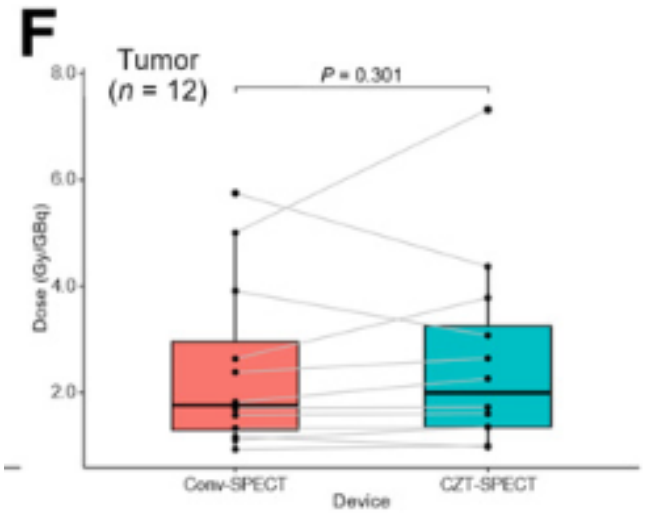


Predicted vs observed

"It is worth noting that, in our study, the predicted absorbed doses tend to underestimate the therapy-delivered absorbed dose when the therapy-delivered absorbed dose was high and overestimate when it was low "

Bland-Altman plots of relative percent error in model prediction versus therapy-delivered absorbed dose with model predictions provided by (A) univariable model with PET uptake and (B) univariable model with eGFR. Horizontal axis is the therapy-delivered renal absorbed dose, and the vertical axis is the relative percent difference between model predictions and delivered dose

Case 2 – Agreement



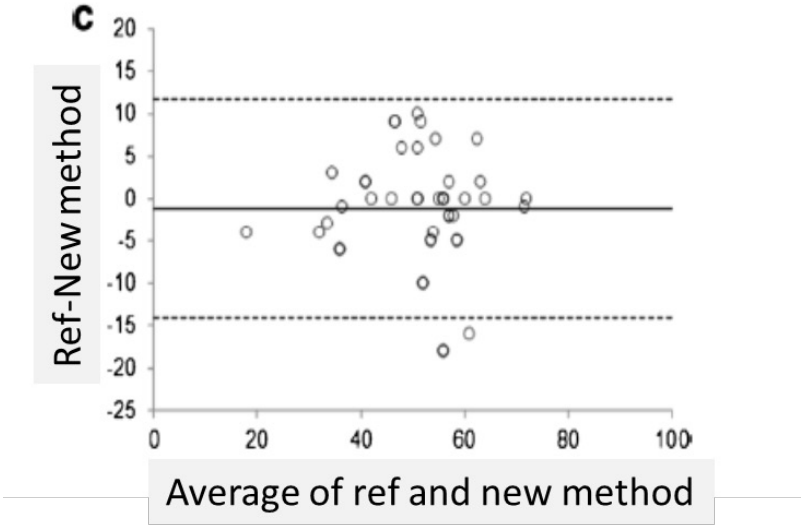
Two methods of clinical measurement

"**Methods**: The mean doses to the organ were compared between the two methods using paired Wilcoxon test for differences

Results: ADs were not significantly different between conventional and NEW Method in most of organs and the tumour

Conclusion: AD calculated with the NEW method are globally comparable to those obtained from a REFERENCE Method

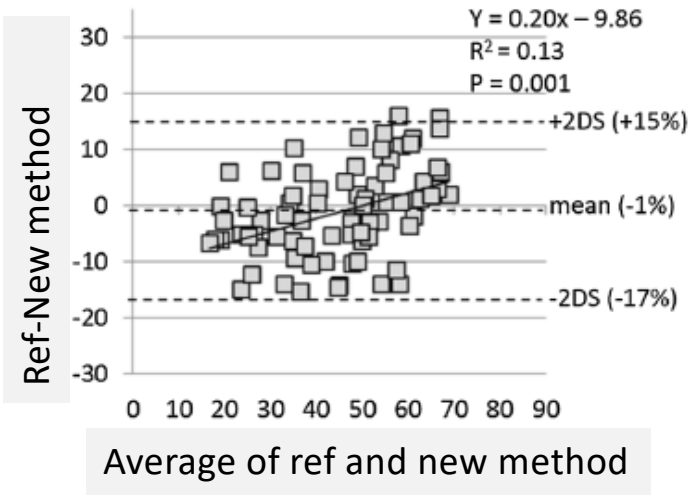
Case 3 – Agreement



Two methods of clinical measurement

"The bias in volume calculation between New method and the Reference Method in the present study was distinctly smaller in amplitude (about 5 %), and therefore unlikely to have been clinically relevant.

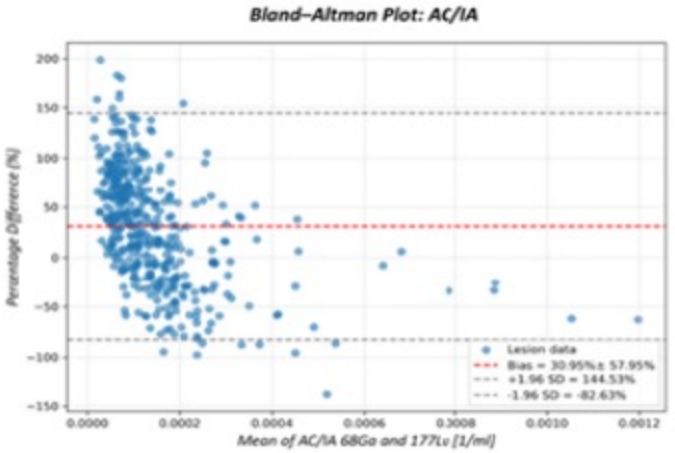
Case 4 – Agreement



Two methods of clinical measurement

A global underestimation was not documented for LV ejection fraction for which the mean difference with Reference Method was -1 ± 8% in the overall population.
However, in a per-patient analysis, this difference was significantly and strongly? related to the level of ejection fraction as shown on the corresponding BA- plot

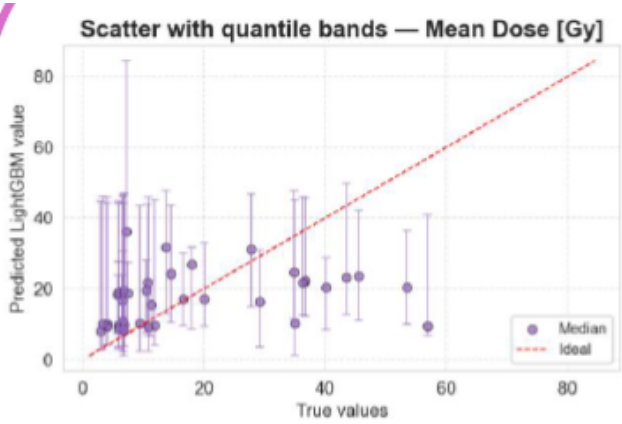
Case 5 – Agreement



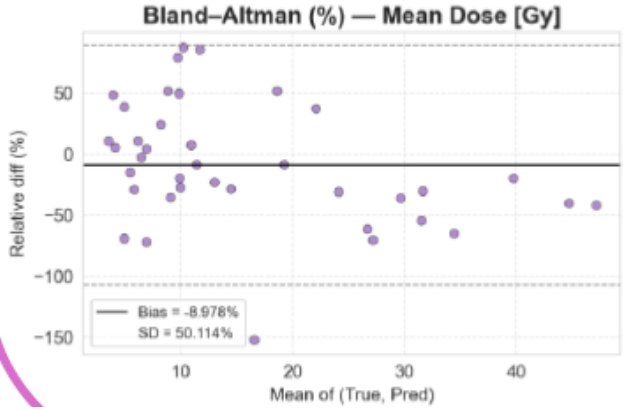
(b) *Bland–Altman plot* (*bias* = +31.0% ±57.95).

Two methods of clinical measurement

Theranostic pair: 68Ga-DOTA-TOC (diagnostic imaging) / 177Lu-DOTATATE (β⁻ therapy + γ post-therapy imaging).

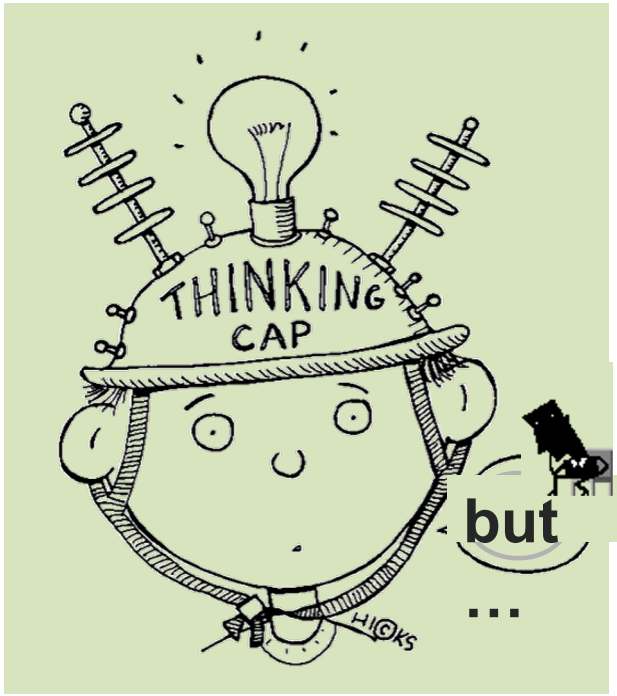


- **Scarse correlation** with higher deviation at higher Mean Doses



- **Little bias ~-9%** of relative %difference between true and predicted values but **HIGH DISPERSION ~50%**

The article that explains the Bland-Altman's method is the **most quoted article** from Lancet magazine, which means that the method has been much quoted or/and applied.



HAS IT BEEN CORRECTLY APPLIED?

Agreement between methods of clinical measurement

The Bland –Altman Method

- In **1980** cardiologists asked for Martin Bland and G. Altman help in assessing the **rate of agreement** between two methods.
- After reviewing the different methods available, they found out none of them served this need and decided to create one of their own.
- **1983** – first publication: The Institute of Statistics
- **1986**: Lancet magazine published the article:
"Statistical methods for assessing agreement between two methods of clinical measurement"
- Most quoted article from Lancet magazine: **> 40 000** citations
- Number 63 among the 100 most highly cited papers of all time [*]

Bland J M and Altman D G

* Richard Van Noorden **These are the most-cited research papers of all time. Nature News. 2025**

Bias and Limits of Agreement

1st Step.

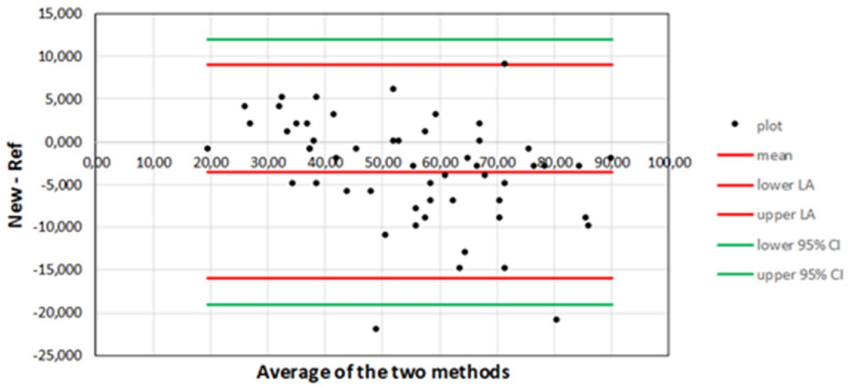
Plot of the difference between the methods against their mean

2nd Step.

Calculate:

- the **bias**, estimated by the mean difference d
- the standard deviation of the differences (*s*)
- the **limits of agreement** (LA)

		95% confidence interval	
mean of differences	-3,519	-5,241	-1,796
standard deviation of differences	6,371		
Lower limit of agreement	-16,006	-18,990	-13,023
Upper limit of agreement	8,969	5,986	11,953



LoA = [d – 2sd ; d + 2sd]

LA – limits of agreement

d – average of differences

sd - standard deviation of difference

It is expected that most of the differences (95%) lie between the limits of agreement.

Bland Altman in steps

1. Define a Medically accepted limit of agreement

You should be able to define a medically accepted limit before conducting the assessment of agreement

Example:

The current guidelines for treating patients with cardiotoxic chemotherapy states that chemotherapy should be considered discontinued if the patient presents again with a drop in LVEF of 10% points or more.

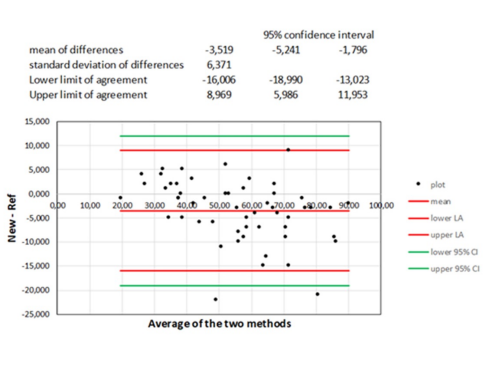
2. Establish the sample size needed for assessing the agreement by BA Method

Sample size depend on α, β, the mean and the SD of differences between two measurements , and the predefined limits

Bland Altman in steps

3. Check the conditions of validity of the BA method

- no relation between the difference and the mean
- the SD is constant
- differences follow a normal distribution



4. Establish the width of the Limits of agreement band

- Upper 95% CI of upper LA –Lower 95% CI of lower LA

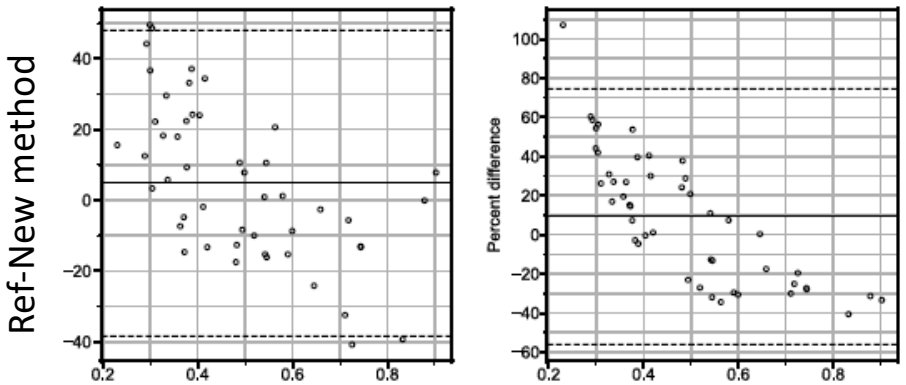
5. Compare 3 with 1

the conclusion is made based on the width of the confidence intervals for the LoAs in comparison to predefined clinical agreement limit:

If 3 ≤ 1 the techniques are interchangeable

If 3 >> 1 the techniques are not interchangeable. There is a substantial disagreement between them

Case 1 – Agreement



How to check?

The correlation coefficient for the difference plot data should be determined -- it should approximate to zero.

Conclusions

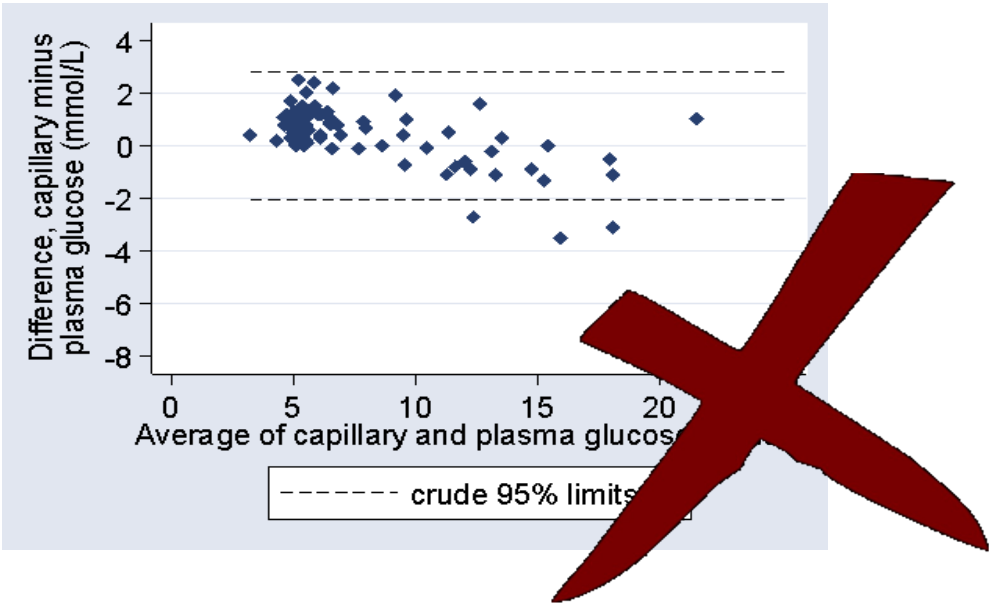
- BA method cannot be applied here
- Use instead the regression diagnostics Predicted vs observed

Bland-Altman's Method

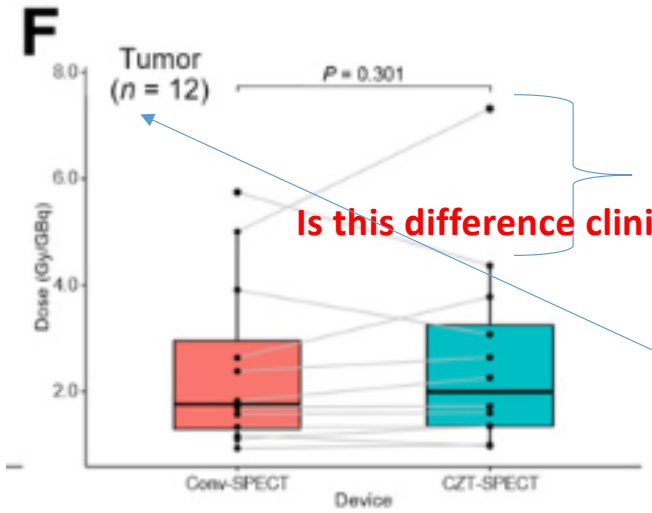
1st Assumption

Make sure that there is no relation between the difference and the mean

Example showing relation between averages and differences



Case 2 – Agreement



Is this difference clinically relevant?

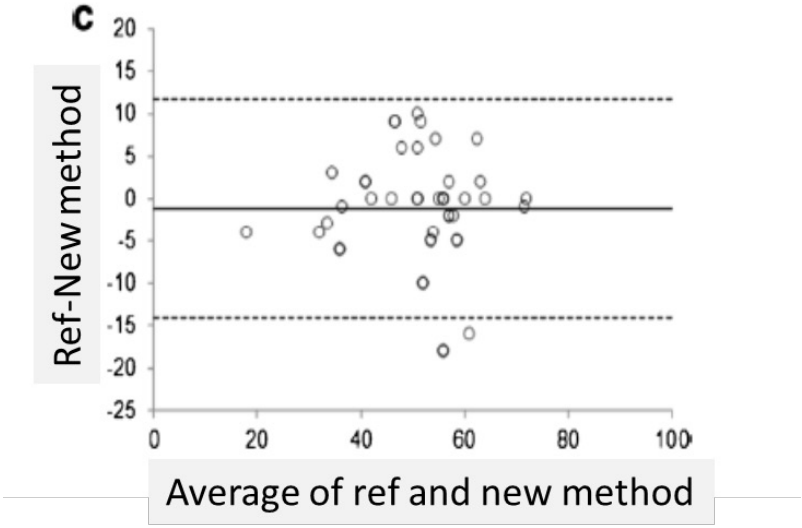
"The mean doses to the organ were compared between the two methods using paired Wilcoxon test for differences"

ADs were not significantly different between conventional and NEW Method in most of organs and the tumour"

"AD calculated with the NEW method are globally comparable to those obtained from a REFERENCE Method

While a test of the mean difference might show a relationship, it's not the best indicator of agreement, especially when sample sizes are small

Case 3 – Agreement



Two methods of clinical measurement

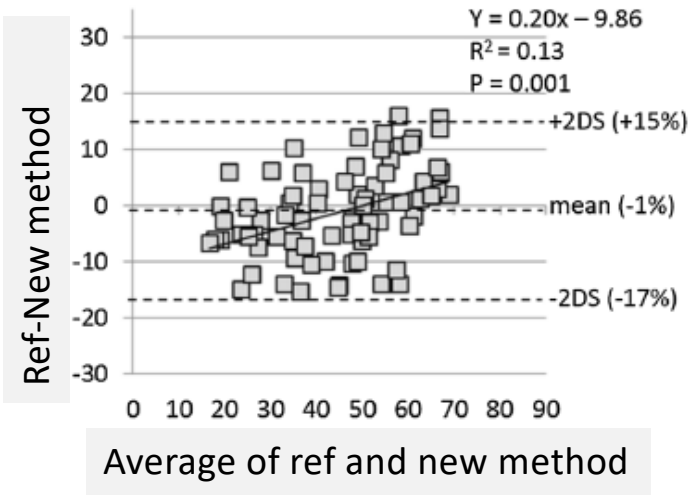
"The bias in volume calculation between New method and the Reference Method in the present study was distinctly smaller in amplitude (about 5 %), and therefore unlikely to have been clinically relevant.

What matters is not the bias but the Limits of agreement (-14 to +12) which are wide demonstrating a poor agreement between the two techniques.

A change of -14% (drop of EF from 50% to 36% could be interpreted as a sufficient reason to suspend the Chemotherapy

A change of +12% could switch an impaired LV function (38%) to a normal LV function (50%).

Case 4 – Agreement



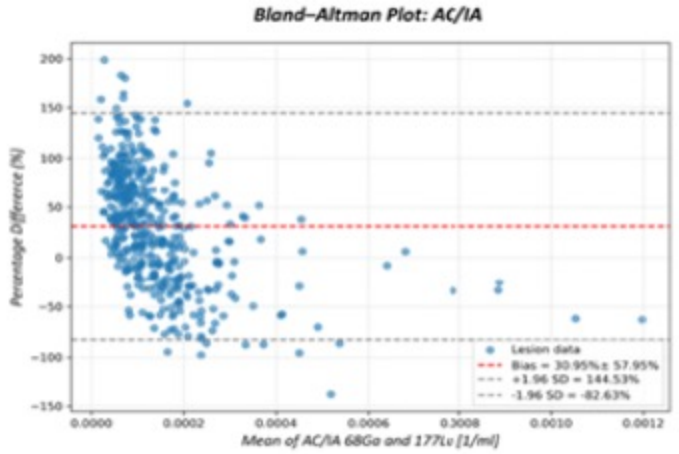
Two methods of clinical measurement

A global underestimation was not documented for LV ejection fraction for which the mean difference with Reference Method was -1 ± 8% in the overall population.
However, in a per-patient analysis, this difference was significantly and strongly? related to the level of ejection fraction as shown on the corresponding BA- plot

What matters is not the bias but the Limits of agreement (-17 to +15) which are wide demonstrating a poor agreement between the two techniques.
A change of -17 (drop of EF from 50% to 33% could be interpreted as a sufficient reason to suspend the Chemotherapy
A change of +15 could switch an impaired LV function (35%) to a normal LV function (50%).

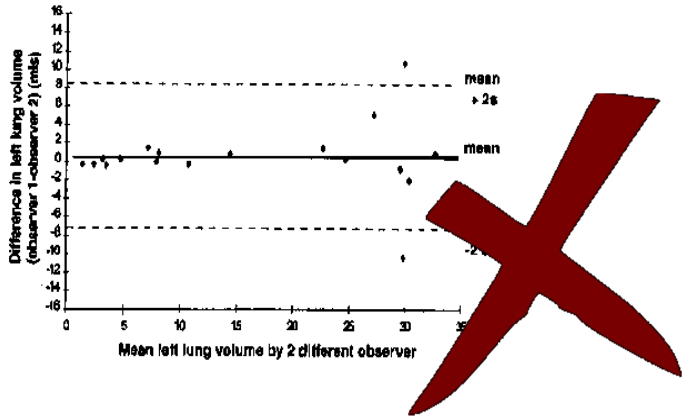
An R^2=0.13 cannot be defined as a strong correlation! Only 13% of the variance is explained by the relationship.

Case 5 – absolute vs percentage differences



(b) *Bland–Altman plot* (*bias* = +31.0% ±57.95).

Example showing not constant *sd* (standard deviation)



Use Bland-Altman **absolute difference** when the measurement units are consistent and you want to know the raw error between methods. Use the **percentage difference** when the variability of the difference increases with the magnitude of the measurement, as this normalizes the difference to the value itself.

Feature	Bland-Altman Absolute Difference	Bland-Altman Percentage Difference
Purpose	Compares two methods by plotting the raw difference against the average of the two measurements.	Compares two methods by plotting the difference as a percentage of the average measurement.
When to use	When the variability is relatively constant across the range of measurements.	When the variability in the difference increases as the magnitude of the measurement increases (heteroscedasticity).
How it works	The y-axis shows the direct difference (Method A - Method B).	The y-axis shows the percentage difference: `((Method A - Method B) / average) * 100`.
Key takeaway	Shows the average bias and the limits of agreement in the same units as the original data.	Shows the average bias and limits of agreement as a proportion of the measurement value, making it useful for comparing relative error.

Lin's Concordance Correlation Coefficient

Like a correlation, CCC ρ_c ranges from -1 to 1, with perfect agreement at 1. It cannot exceed the absolute value of ρ, Pearson's correlation coefficient between Y and X. It can be legitimately calculated on as few as ten observations. Following Lin et al. (2002),

$$\rho_C = \frac{2 \cdot \rho \cdot \sigma_{ref} \cdot \sigma_s}{\sigma_{ref}^2 + \sigma_s^2 + (\mu_{ref} - \mu_s)^2}$$

With:

ρ the correlation coefficient between the reference standard and the new method

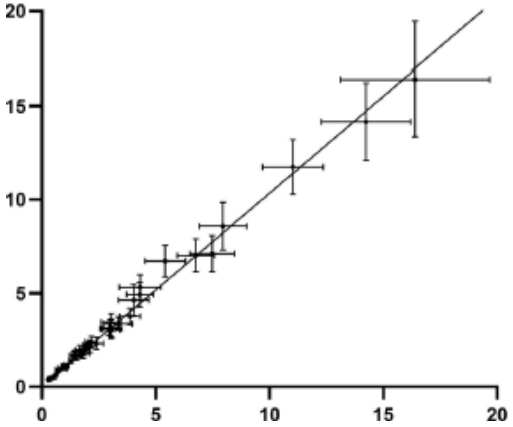
σ_{ref} and σ_s are the variance of the reference standard and the new method

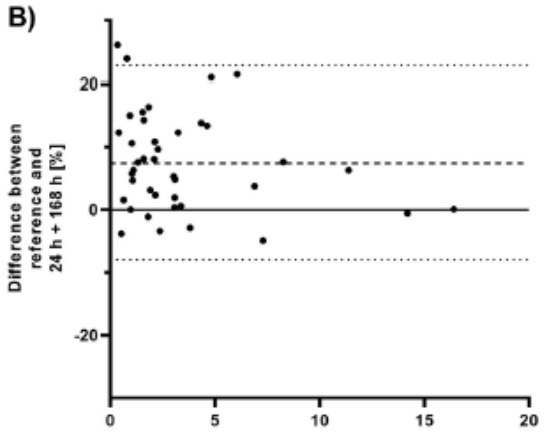
μ_{ref} and μ_s are the means of the reference standard and the new method

Strength-of-agreement	Continuous variables
Almost perfect	>0.99
Substantial	0.95..0.99
Moderate	0.90..0.95
Poor	<0.90

To assess the degree of agreement for a given set of data it is proposed that the lower one side 95% confidence limit for the calculated concordance correlation coefficient should be compared to the values in this table.

It is desirable that the assessment be performed on at least 25 samples, preferably 50.





Method	$u(D)$ (%)	ρ_C	E_N
1 h	**24.3 ± 1.1**	0.79	3.65 ± 1.86
24 h	**21.2 ± 1.2**	**0.96**	1.82 ± 1.32
48 h	**17.5 ± 2.3**	**0.94**	2.30 ± 1.96
72 h	**15.8 ± 1.6**	**0.94**	1.60 ± 1.79
168 h	**20.1 ± 1.3**	**0.98**	1.21 ± 0.94
1 h + 24 h	**20.3 ± 16.1**	0.66	7.24 ± 4.75
1 h + 48 h	**18.8 ± 11.6**	0.77	5.01 ± 3.69
1 h + 72 h	**16.2 ± 6.2**	0.84	3.60 ± 2.02
1 h + 168 h	***13.8 ± 2.0***	***0.99***	***0.68 ± 0.61***
24 h + 48 h	31.2 ± 24.9	0.77	3.69 ± 3.12
24 h + 72 h	**18.7 ± 7.7**	**0.91**	2.27 ± 1.97
24 h + 168 h	***12.9 ± 2.3***	***0.99***	***0.65 ± 0.50***
48 h + 72 h	**24.7 ± 18.1**	**0.92**	2.79 ± 1.89
48 h + 168 h	***13.1 ± 3.0***	***0.99***	***0.65 ± 0.64***
72 h + 168 h	***13.4 ± 2.4***	***0.99***	***0.86 ± 0.48***

Selection based on the value of CCC (>0.9) and uncertainty (%)

Conclusions

Pearson’s correlation coefficient

high correlation does not mean that the two methods agree. Indeed, r measures the strength of a relation between two variables, not the agreement between them and data which seem to be in poor agreement can produce quite high correlations.

Tests of the significance of the correlation value

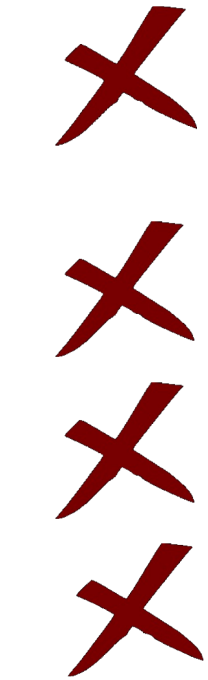
the test of significance of the correlation value is irrelevant to the question of agreement. Indeed, it would be amazing if two methods designed to measure the same quantity were not related.

Tests of the significance between mean values

While a test of the mean difference might show a relationship, it's not the best indicator of agreement, especially when sample sizes are small

K statistic or intraclass correlation coefficients

These are indexes of reproducibility, not of agreement.



Bland-Altman Method

Correct. What matter most are the LAs not the bias. Consider the 95% CI of the Las

Lin’s concordance correlation coefficient

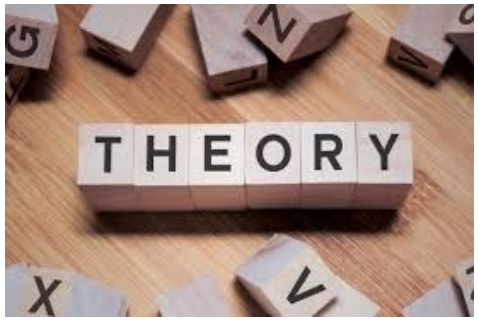
Correct. It must be used in conjunction with strength-of agreement criteria

Correlation, Regression

(dose –response relationship)

- Explanation: R^2, standardized regression coefficients
- Prediction: Regression, unstandardized regression coefficients

1. θεωρία



2. παραδείγματα

Assumptions of linear regression

For linear models, the dependent variable doesn't have to be normally distributed, but it does have to be continuous, unbounded, and measured on an interval or ratio scale.

1. Linearity (a linear relationship between dependent and independent variables)
2. Independence (error terms are independent of each other)
    1. the assumption of independence refers to the residual errors being independent of each other, not the independent variables. This means knowing the error for one data point should provide no information about the error for another, which is particularly important for time-series data and is also called autocorrelation
3. Normality (error terms are normally distributed)
    1. While independent variables don't need to be normally distributed, the residuals should ideally follow a normal distribution for the most accurate p-values and confidence intervals in significance testing.
4. Homoscedasticity (error terms have constant variance)

- Additionally, for multiple regression, the assumption of no multicollinearity (independent variables are not highly correlated) is critical.

Validity of assumptions is seldom checked and reported!!!

Common mistakes in linear regression

Data and model specification errors

- Omitted variable bias
- Non linear relationship
- Outliers
- Autocorrelation
- Overfitting

Interpretation and validity errors

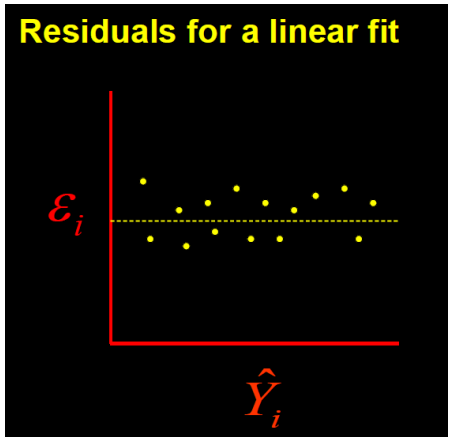
- Multi collinearity
- Misinterpreting coefficients
- Correlation vs causation
- Simultaneous causality
- Extrapolation

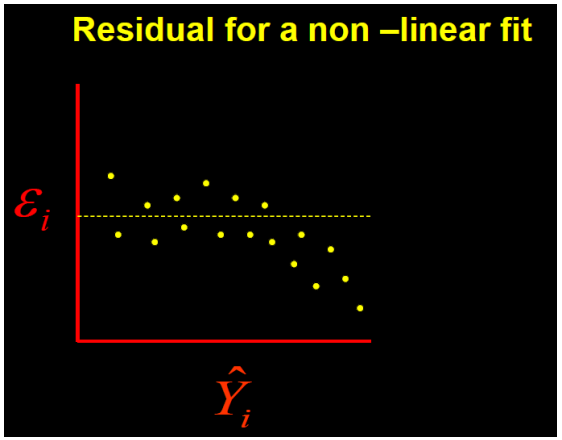
Assumptions violations

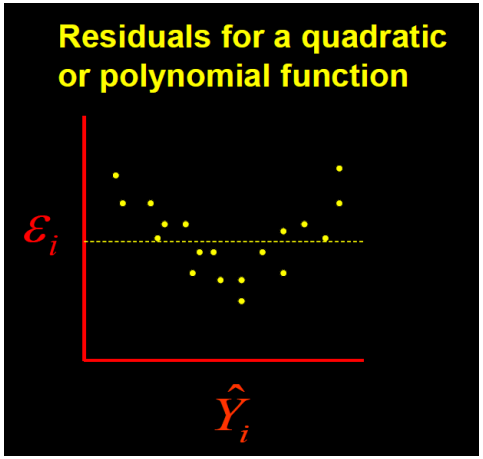
- Non-normal errors
- Homoscedasticity
- Independence of errors
- Measurement errors

Check of Assumptions of linear regression

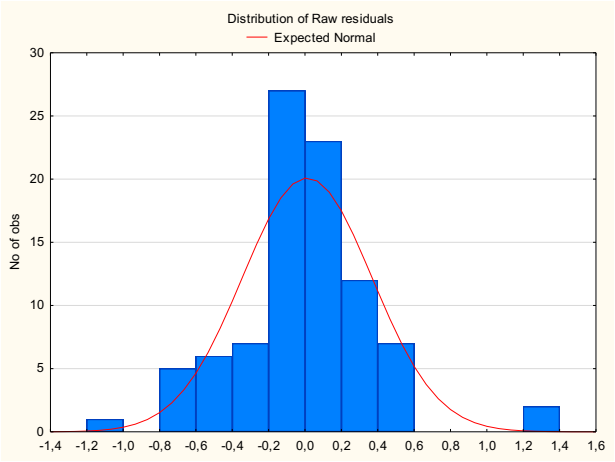
Linearity

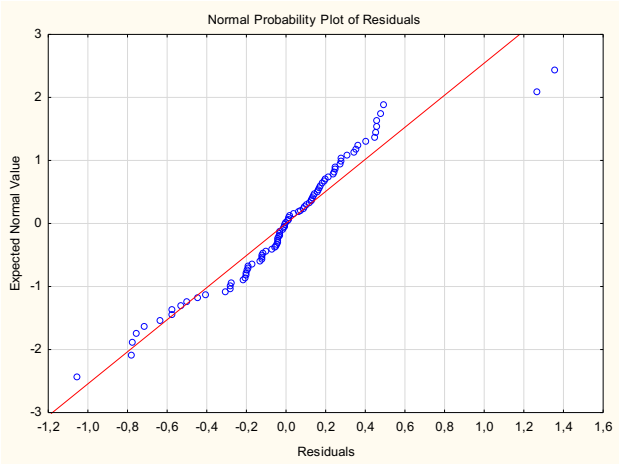


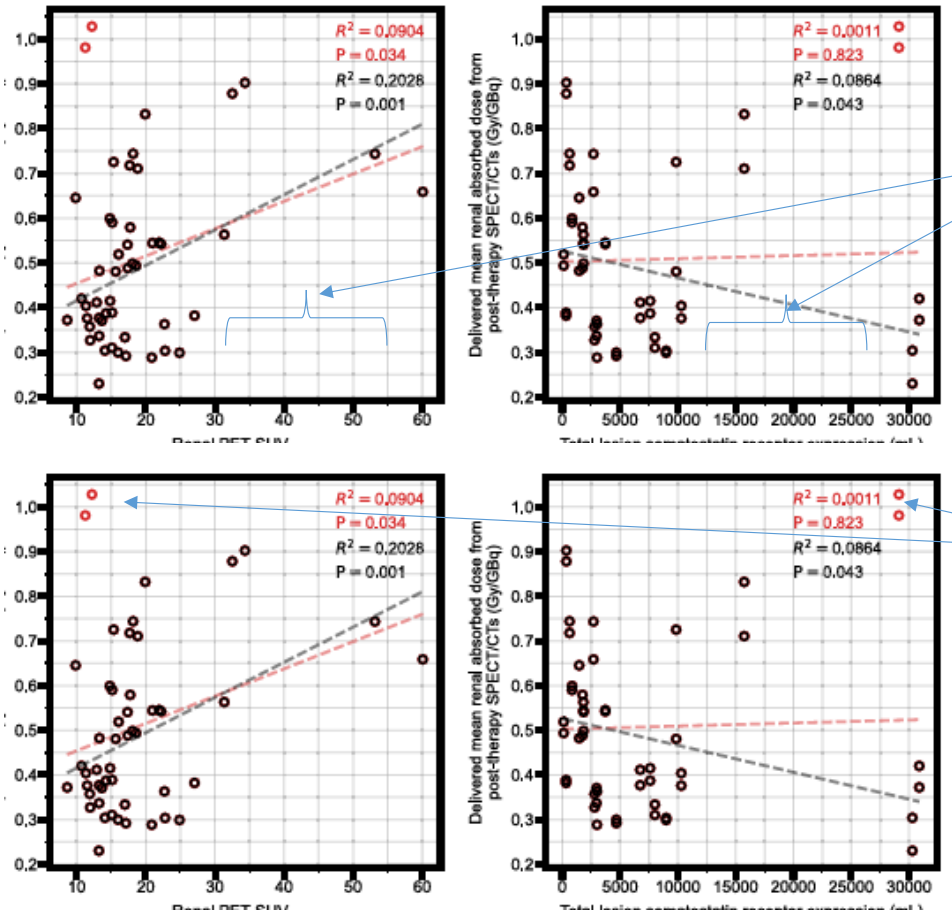




Normality of residuals







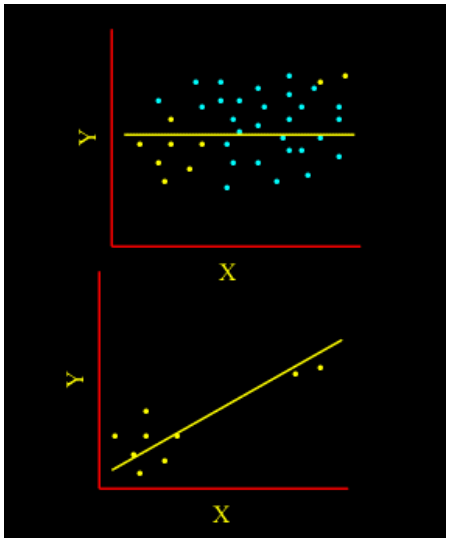
Regression – important points

1. Ensure that the distribution of the values of the prediction variables are approximately uniform within the sampled range. **Extrapolation**

1. p value is irrelevant (p<0.05 but there is no clinically significant correlation)

2. Manage carefully **Outliers**!

3. The exclusion of only one patient modify completely the regression patterns. Does this tell us something about the robustness of the estimates?



5. If you do not adopt a split sample approach you cannot test prediction accuracy. Only explanation.

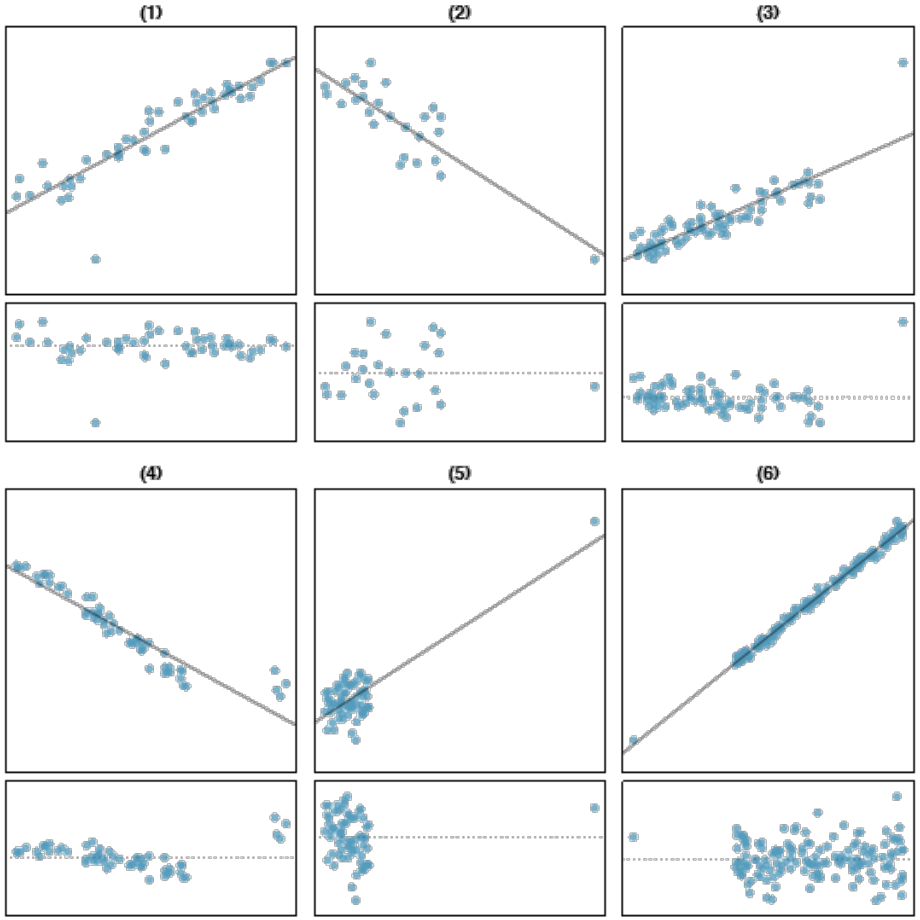
6. An R^2 of 0.20 means that you are explaining only 20% of the variance of the dependent variable. In other words 80% of the variance is not explained by the model

Types of Outliers in Linear Regression

There are six plots shown in Figure along with the least squares line and residual plots. For each scatter plot and residual plot pair, identify any obvious outliers and note how they influence the least squares line. Recall that an outlier is any point that doesn't appear to belong with the vast majority of the other points.

1.There is one outlier far from the other points, though it only appears to slightly influence the line.
2.There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.
3.There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to t very well.
4.There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least square line t poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.
5.There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.
6.There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential.

Examine the residual plots in Figure . You will probably find that there is some trend in the main clouds of (3) and (4). In these cases, the outliers influenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

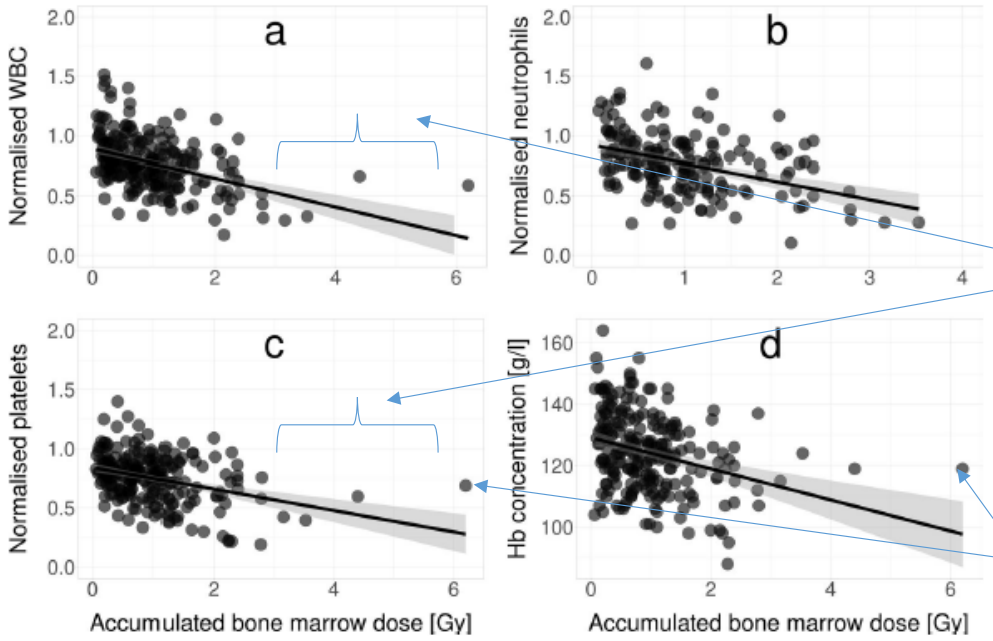


It is tempting to remove outliers. Do not do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly.

https://www.openintro.org/

$r = -0.43$,

$r = -0.42$,

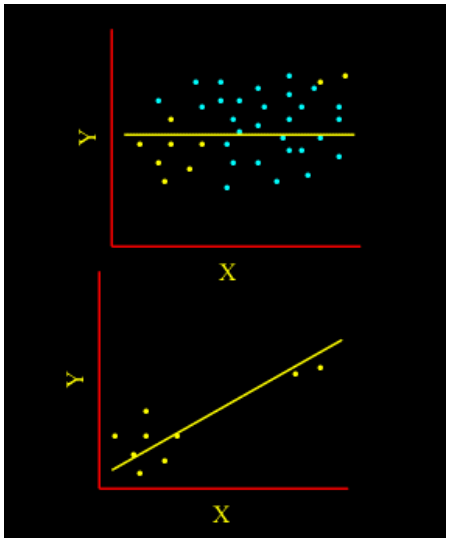


$r = -0.35$,

$r = -0.31$,

Regression – important points

1. Ensure that the distribution of the values of the prediction variables are approximately uniform within the sampled range. **Extrapolation**



2. p value is irrelevant (p<0.05 but there is no clinically significant correlation)

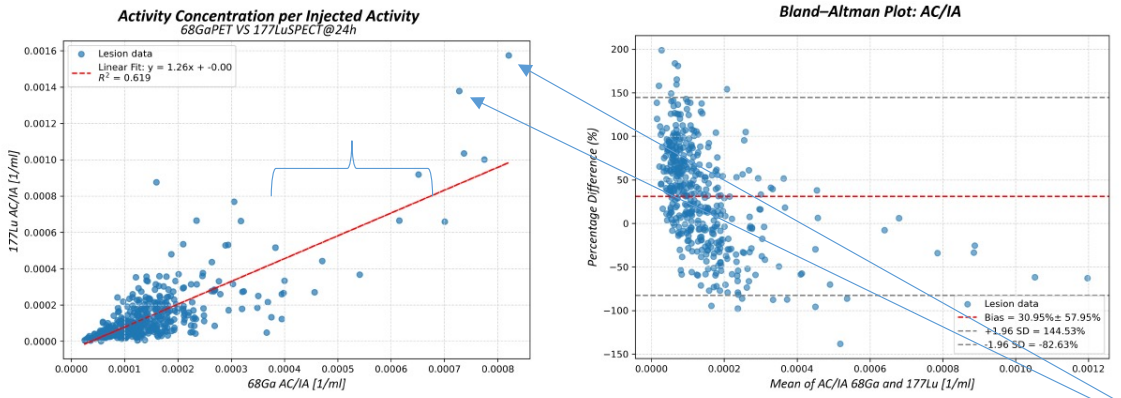
3. Manage carefully **Outliers**!

4. Don't forget about **Autocorrelation** *(In the resulting dataset, data from 940 blood sample time points in connection to 229 treatments in 59 patients were included).*

5. If you do not adopt a split sample approach you cannot do prediction. Only explanation.

6. An r=-0.31 means that you are explaining only 9% of the variance of the dependent variable. In other words 91% of the variance is not explained by the model

Regression – important points

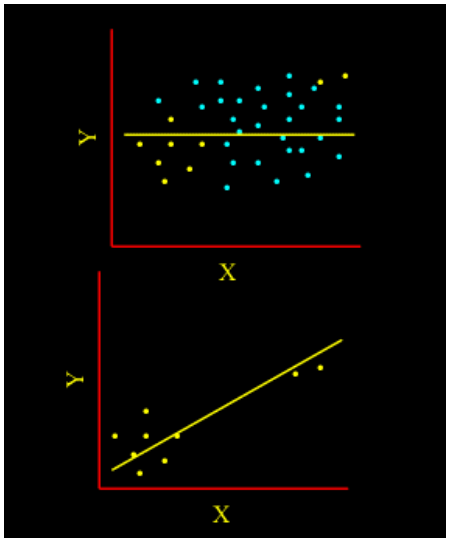


(a) *Scatter plot* (AC/IA, $R^2 = 0.619$). (b) *Bland–Altman plot* ($bias = +31.0\% \pm 57.95$).

Figure 5.8: *Correlation of the entire dataset between PET- and SPECT-derived normalized activity concentration* (AC/IA). *Moderate correlation with positive bias and a large SD..*

1. Ensure that the distribution of the values of the prediction variables are approximately uniform within the sampled range. **Extrapolation**

2. Manage carefully **Outliers**!



3. An r^2=0.61 means that you have an r=0.78
4. While an high r doesn't necessarily mean a high agreement a low r imply necessarily a poor agreement!
5. Prefers absolute changes with respect to % differences

Autocorrelation

In many cases, the value of a variable at a point in time is related to the value of it at a previous point in time. Autocorrelation analysis measures the relationship of the observations between the different points in time, and thus seeks a pattern or trend over the time series. For example, the temperatures on different days in a month are autocorrelated.

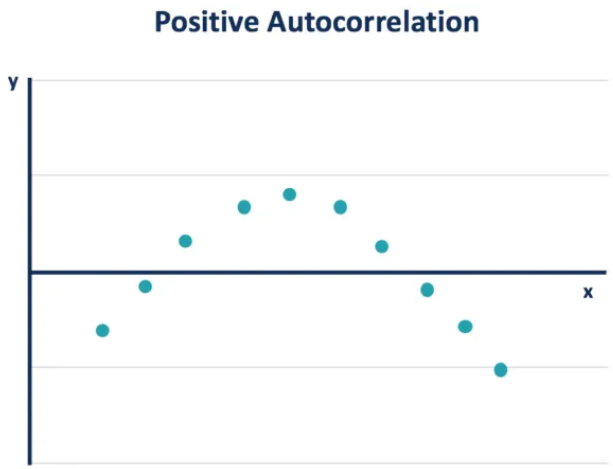
1.Similar to correlation, autocorrelation can be either positive or negative. It ranges from -1 (perfectly negative autocorrelation) to 1 (perfectly positive autocorrelation).

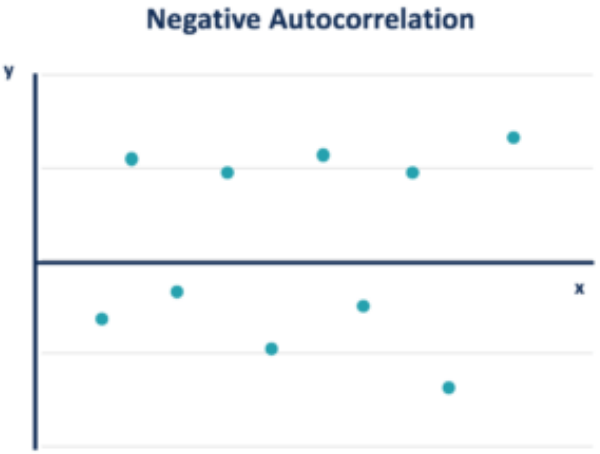
2.Positive autocorrelation means that the increase observed in a time interval leads to a proportionate increase in the lagged time interval.

3.Conversely, negative autocorrelation represents that the increase observed in a time interval leads to a proportionate decrease in the lagged time interval. By plotting the observations with a regression line, it shows that a positive error will be followed by a negative one and vice versa.

4.Test for autocorrelation

The Durbin-Watson statistic is commonly used to test for autocorrelation. It can be applied to a data set by statistical software. The outcome of the Durbin-Watson test ranges from 0 to 4. An outcome closely around 2 means a very low level of autocorrelation. An outcome closer to 0 suggests a stronger positive autocorrelation, and an outcome closer to 4 suggests a stronger negative autocorrelation.





It is necessary to test for autocorrelation when analysing a set of temporal data.

Multi Collinearity

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

Why is Multicollinearity a Potential Problem?

The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you hold all of the other independent variables constant.

However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison.

What Problems Do Multicollinearity Cause?

A high degree of collinearity produces unacceptable uncertainty (large variance) in the estimates of the regression coefficients (i.e., a large sampling variation).

Specifically, the coefficients can change dramatically depending on which terms are included or not in the model and also on the order in which they are placed in the model.

It does not affect the prediction (the predicted values), but it does affect the interpretation of the slopes (contribution of the variables).

Applied Linear Statistical Models, p289, 4th Edition.

Multi Collinearity

Do I Have to Fix Multicollinearity?

The need to reduce Multicollinearity depends on its severity and your primary goal for your regression model. Keep the following three points in mind:

1.The severity of the problems increases with the degree of the Multicollinearity. Therefore, if you have only moderate Multicollinearity, you may not need to resolve it.

2.Multicollinearity affects only the specific independent variables that are correlated. Therefore, if Multicollinearity is not present for the independent variables that you are particularly interested in, you may not need to resolve it. Suppose your model contains the experimental variables of interest and some control variables. If high Multicollinearity exists for the control variables but not the experimental variables, then you can interpret the experimental variables without problems.

3.Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don’t need to understand the role of each independent variable, you don’t need to reduce severe Multicollinearity.

Multi Collinearity

How to detect Multicollinearity?

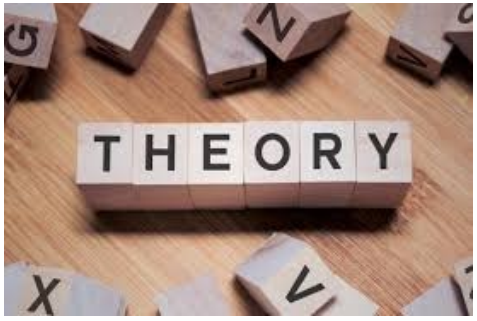
Variance inflation factor (VIF) and tolerance are two related statistics used to diagnose Multicollinearity, or high correlation between predictor variables, in a multiple regression model. Tolerance is the reciprocal of VIF, meaning Tolerance = 1/VIF1 .

A VIF greater than 10, or a tolerance less than 0.10, suggests problematic Multicollinearity.

Dose-Response relationship

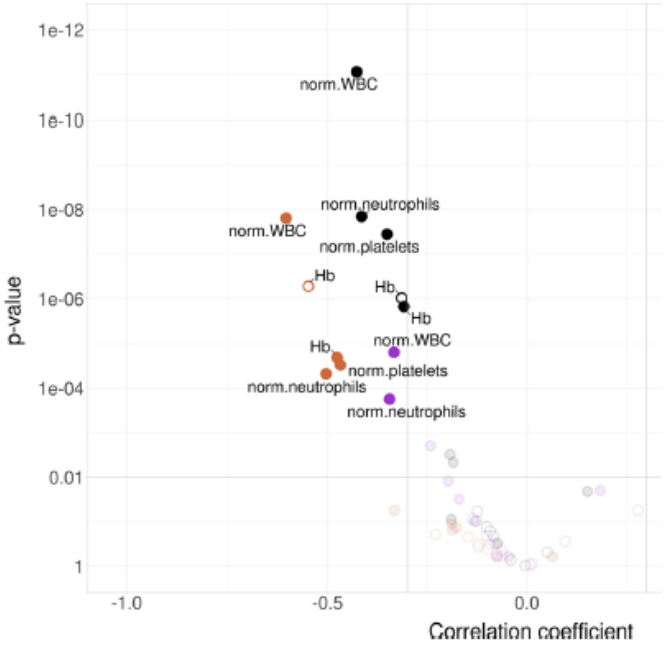
1. παραδείγματα

2. θεωρία

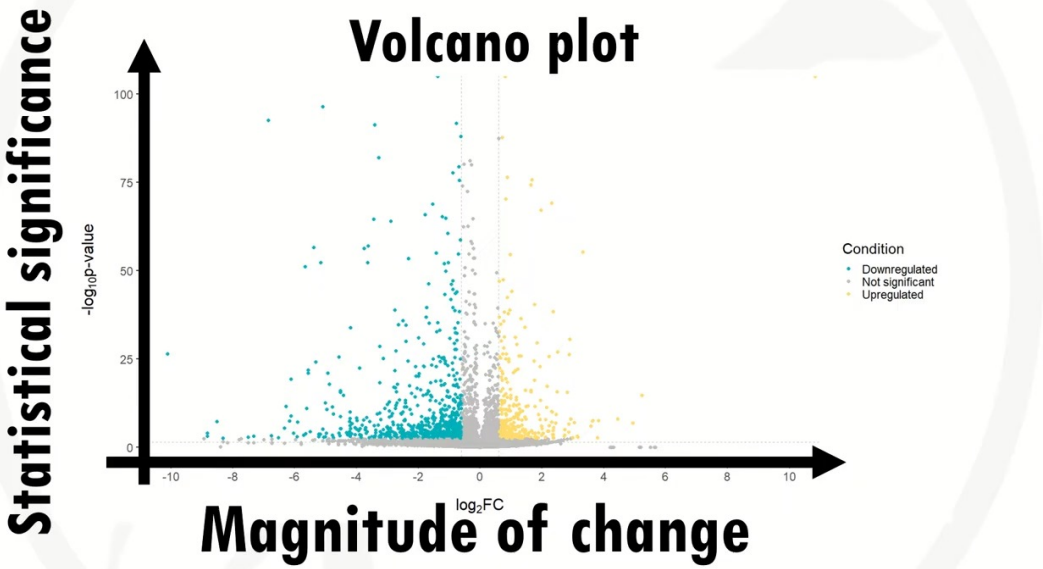


NOT a Dose –response relationship

In total, 69patients were eligible.



Dose–response relationships Dose–response relationships between ……



A volcano plot is a type of scatter plot used to visualize differential expression in large biological datasets, such as gene or protein expression, by showing statistical significance (on the y-axis) versus fold change (on the x-axis) for each data point.

p value is irrelevant. p can be <<0.05 but there still can't be any clinically significant correlation!
This graph cannot by any mean be considered as showing a dose-response relationship

How to establish a significant difference among groups? Temporal Trends

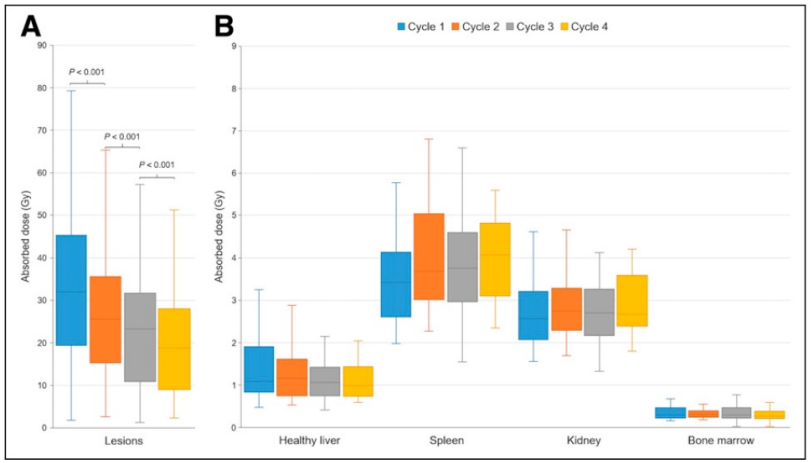


FIGURE 1. Distribution of ADs by lesions (A) and selected healthy organs (B) in 4 PPRT cycles.

Statistical analysis

The paired Student t-test was used to detect differences between mean values in the same population or the same lesion.

The ADs by healthy organs were not significantly different among [177Lu]Lu-DOTATATE cycles except for the spleen ($P < 0.05$).

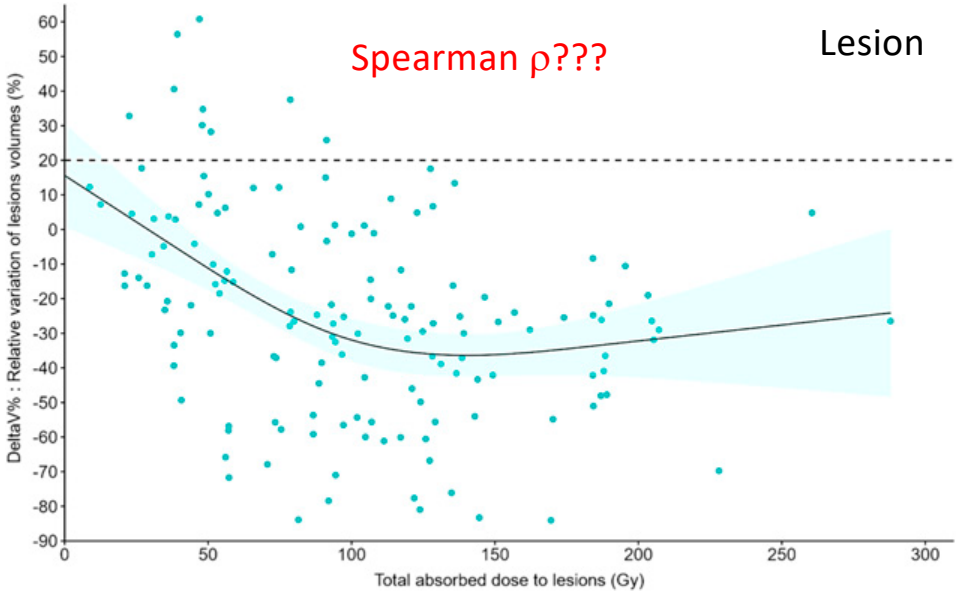
The ADs by lesions decreased significantly over time from cycle 1 to cycle 4 ($P < 0.001$).

A repeated measures ANOVA is a statistical test used to compare the means at three or more time points where the same participants are measured multiple times.

How it works

- Unlike a traditional one-way ANOVA, which compares independent groups, repeated measures ANOVA uses the same subjects for each group or time point.
- This design is statistically powerful because each participant serves as their own control, which helps to reduce the impact of individual differences.
- The test analyzes whether there is a statistically significant difference between the means of these related groups.

How to fit the data?



LOESS (locally estimated scatterplot smoothing)

How it works

1. Basically it is a locally adjusted quadratic fit
2. Consider the fraction Alfa (span) of points in the data set that are nearest to X
3. Weight points close to X more than points farter away
4. Fit a quadratic model using that weighted data set
5. Use that model to predict the value of the response variable for the explanatory value X

Pro and Cons

Loess models are very flexible. They are non parametric and can fit any distribution of data. There are disadvantages as well, however.

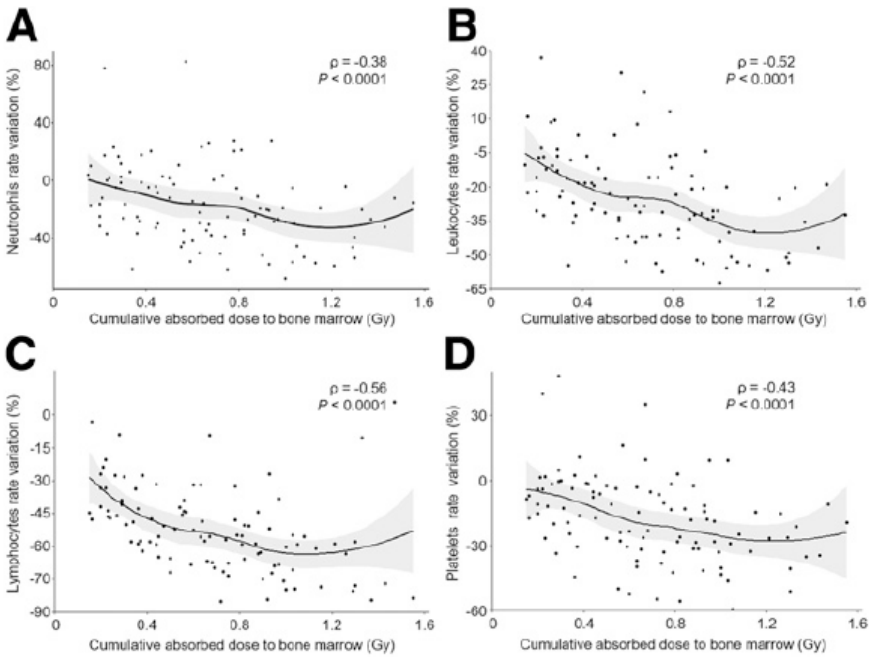
They are not transparent or easy to interpret. This is particularly true when multiple explanatory variables are present, which is not the case here.

They are prone to overfitting

The hyper parameter alfa (span) must be tuned, which requires knowledge and care.

How to interpret the relationships?

OARs



Conclusion:

In patients treated with [$_{177}$Lu]Lu-DOTATATE for GEP-NETs, tumour and healthy organ dosimetry can predict survival and toxicities, thus influencing clinical management.

Comments:

With ρ of 0.4 -0.5 you might expect to explain roughly 20% 25% of variability in the data.

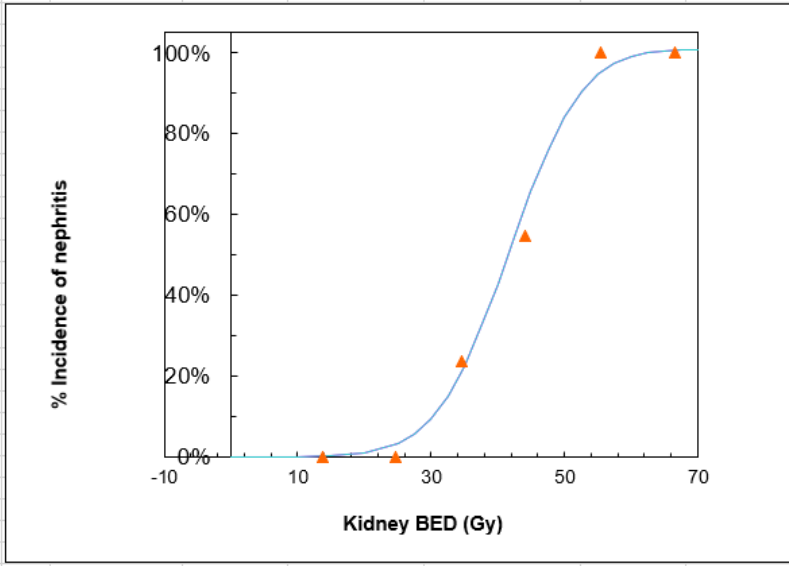
Consider rewording the sentence as:

“tumour and healthy organ dosimetry could partially explain survival and toxicities, thus influencing clinical management”

Building NTCP curves

				threshold for toxicity:	10
TD50	43		Patient	BED (Gy)	CLR (%/year)	CLR: Y=1, N=0
m	0,2		1	16,1	0	0
BED (Gy)	t	NTCP Barone+IEO	2	17,7	0	0
0,01	-5,00	0,00	3	11,0	0	0
1	-4,88	0,00	4	10,5	0	0
2	-4,77	0,00	5	27,7	1,4	0
5	-4,42	0,00	6	28,1	3,1	0
10	-3,84	0,00	7	21,3	0	0
15	-3,26	0,00	8	22,0	0	0
20	-2,67	0,01	9	38,1	14	1
25	-2,09	0,03	10	33,8	0	0
27,5	-1,80	0,06	11	30,2	2,8	0
30	-1,51	0,09	12	36,3	9,9	0
32,5	-1,22	0,15	13	31,0	2	0
35	-0,93	0,22	14	35,5	0	0
37,5	-0,64	0,32	15	31,5	4,4	0
40	-0,35	0,43	16	33,1	0,9	0
42,5	-0,06	0,54	17	33,7	16	1
45	0,23	0,66	18	33,2	0	0
47,5	0,52	0,76	19	39,2	32	1
50	0,81	0,84	20	36,6	0	0
52,5	1,10	0,90	21	38,7	0	0
55	1,40	0,95	22	37,5	20	1
57,5	1,69	0,98	23	31,1	0	0
60	1,98	0,99	24	38,9	0	0
62,5	2,27	1,00	25	30,5	0	0
65	2,56	1,01	26	45,8	23	1
67,5	2,85	1,01	27	47,0	40	1
70	3,14	1,01	28	49,4	4	0
72,5	3,43	1,01	29	47,4	30	1
75	3,72	1,01	30	43,5	20	1
77,5	4,01	1,01	31	42,9	21	1
80	4,30	1,01	32	41,1	39	1
82,5	4,59	1,01	33	44,4	0	0
85	4,88	1,01	34	40,5	4,2	0
87,5	5,17	1,01	35	41	9,4	0
90	5,47	1,01	36	42,7	5,9	0
92,5	5,76	1,01	37	55,2	53	1
			38	54,8	26	1
			39	56	38,8	1
			40	56,1	56,4	1
			41	50,8	45,7	1
			42	59,3	51,3	1
			43	66,5	13	1

		CLR > 10%			
Absorbed ose (Gy)	# patients	# cases w. toxicity	BED mean (Gy)		
20	0	0	14	0%	0%
30	0	0	25	0%	1%
40	17	4	35	24%	15%
50	11	6	44	55%	54%
60	6	6	55	100%	95%
70	0	1	67	100%	101%
total	34				24%



Building NTCP curves

General Linear Models

- Family of regression models
- Response — Model Type
  - Continuous — Linear regression
  - Counts — Poisson regression
  - Survival times — Cox model
  - Binomial — Logistic regression
- Uses
  - Control for potentially confounding factors
  - Model building , risk prediction

Logistic Regression

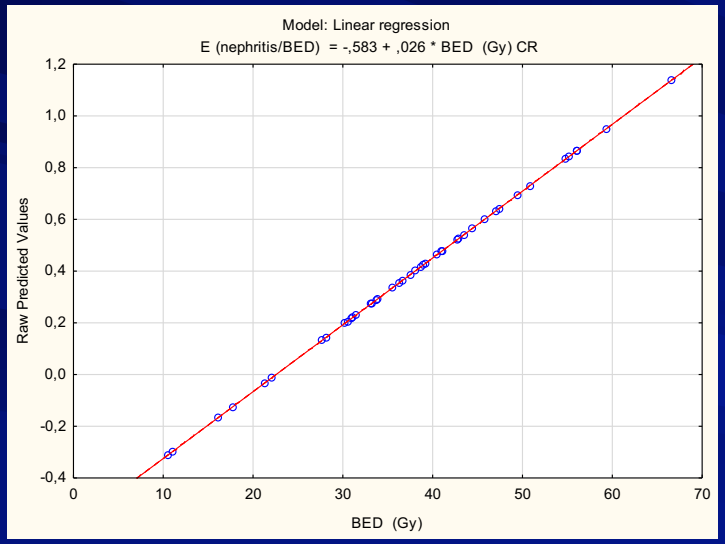
- Models relationship between set of variables X_i
  - dichotomous (yes/no, smoker/nonsmoker,…)
  - categorical (social class, race, ...)
  - continuous (age, weight, gestational age, ...)

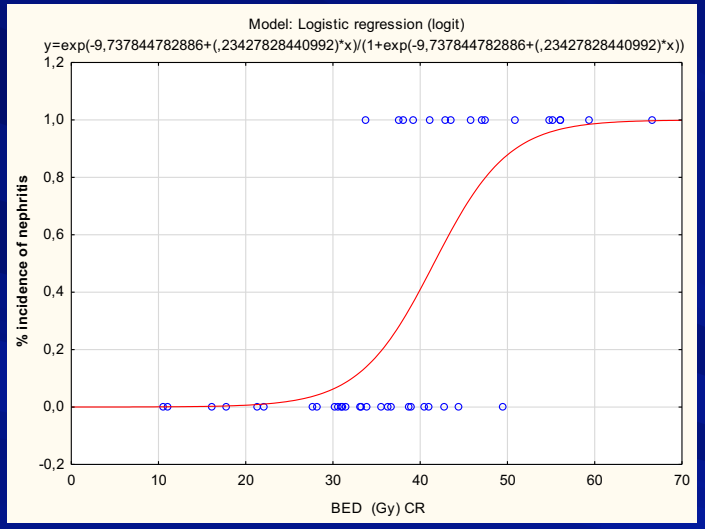
and

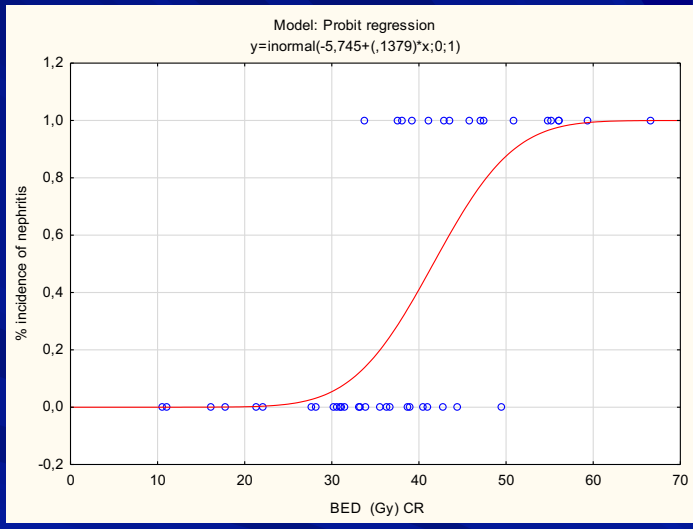
  - dichotomous categorical response variable Y

    e.g. Success/Failure, Remission/No Remission Survived/Died, CHD/No CHD, Low Birth Weight/Normal Birth Weight, etc…

Logistic Regression







$$E(nephritis \mid BED) = -.583 + .026 \cdot BED(Gy)$$

e.g. For an individual with BED = 50 Gy

$$E(nephritis \mid BED = 50) = -.583 + .026 \cdot 50 = .72\,??$$

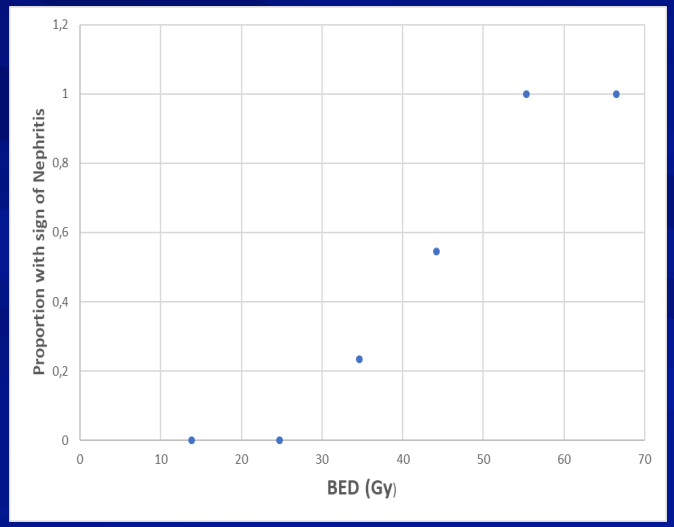
The smooth regression estimate is "S-shaped" but what does the estimated mean value represent?

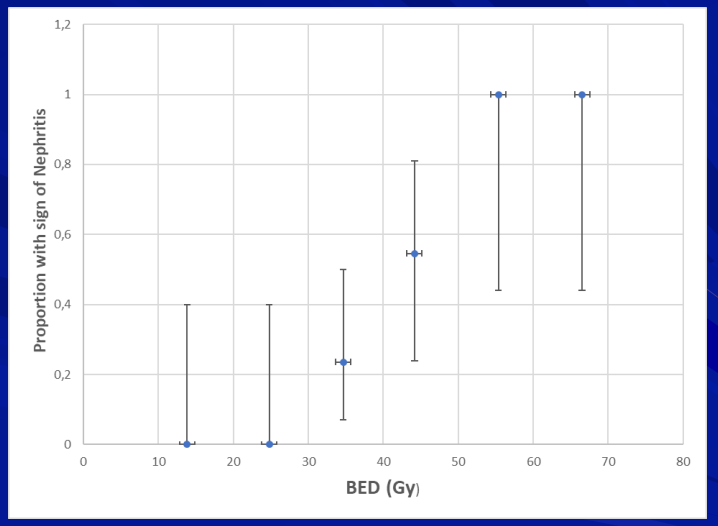
Answer: P(Nephritis|BED)!!!!

Logistic Regression

We can group individuals into BED and look at the percentage/proportion showing signs of nephritis

		Diseased	
Age group	# in group	#	Proportion
1) 20-29	4	0	0
2) 30-39	4	0	0
3) 40-49	17	4	0,24 (0.07-0.50)
4) 50-59	11	6	0,55 (0.24-0.81)
5) 60-69	6	6	1
6) >70	1	1	1



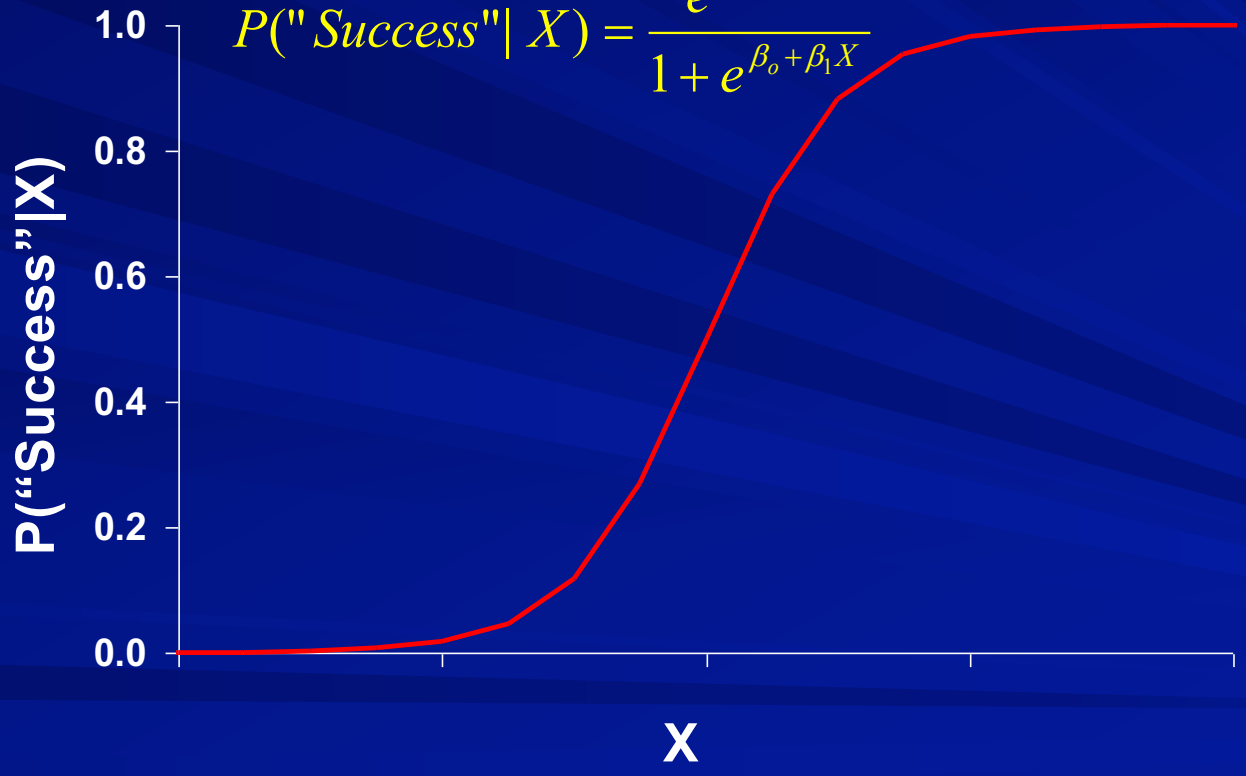


Notice the "S-shape" to the estimated proportions vs. BED.

Notice the wide confidence intervals which are a warning against over interpretation of the estimates.

Logistic Function

$$P(\text{"}Success\text{"}|\ X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$



Logit Transformation

The logistic regression model is given by

$$P(Y \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

which is equivalent to

$$\underbrace{\ln\left(\frac{P(Y \mid X)}{1 - P(Y \mid X)}\right)} = \beta_o + \beta_1 X$$

This is called the Logit Transformation

Dichotomous Predictor

Consider a dichotomous predictor (X) which represents the presence of risk (1 = present)

	Risk Factor (X)	
Disease (Y)	**Present** (X = 1)	**Absent** (X = 0)
Yes (Y = 1)	$P(Y=1\|X=1)$	$P(Y=1\|X=0)$
No (Y = 0)	$1-P(Y=1\|X=1)$	$1-P(Y=1\|X=0)$

$$\frac{P}{1-P} = e^{\beta_o+\beta_1 X}$$

$$\text{Odds for Disease with Risk Present} = \frac{P(Y=1\,|\,X=1)}{1-P(Y=1\,|\,X=1)} = e^{\beta_o+\beta_1}$$

$$\text{Odds for Disease with Risk Absent} = \frac{P(Y=1\,|\,X=0)}{1-P(Y=1\,|\,X=0)} = e^{\beta_o}$$

Therefore the odds ratio (OR) $= \dfrac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \dfrac{e^{\beta_o+\beta_1}}{e^{\beta_o}} = e^{\beta_1}$

Dichotomous Predictor

- Therefore, for the odds ratio associated with risk presence we have $OR = e^{\beta_1}$

- Taking the natural logarithm we have

$$\ln(OR) = \beta_1$$

thus the estimated regression coefficient associated with a 0-1 coded dichotomous predictor is the natural log of the OR associated with risk presence!!!

Statistical Problems in deriving NTCP curves

Data and model limitations

•**Simplification of 3D dose distribution:**
•Traditional models rely on dose-volume histograms (DVHs), which condense the entire 3D dose into a 2D representation, ignoring the spatial complexity of the dose distribution.
•**Non-linear relationships:** The relationship between radiation dose and toxicity is often non-linear, and many models do not adequately capture these complex relationships.

Methodological and technical challenges

•**Multicollinearity:** Predictors, especially different dose parameters for the same organ, can be highly correlated, leading to unstable model coefficients.
•**Overfitting:** Models can be overly specific to the training data, leading to poor performance when applied to new patients or cohorts.

Generalizability and validation issues

Lack of generalizability: Models often perform poorly on different patient cohorts or treatment techniques because they fail to account for other contributing factors.
Accounting for other factors: NTCP models often struggle to incorporate other significant factors like concurrent chemotherapy, patient genetics, or organ-specific characteristics, especially for complex organs.
Inter-observer Variation: Differences in how physicians and dosimetrists contour organs on imaging studies can significantly affect the calculated dose-volume parameters, impacting model consistency and reliability

Conclusions

Doctors, by necessity, are becoming the best sceptics in science.

They are the ones forced to weigh weak data against lived patient outcomes, to recognize when a study's "significant" finding is clinically meaningless, and to resist the seduction of novelty.

If researchers embraced that same pragmatism—valuing replication as highly as discovery—perhaps the literature would become as trustworthy as the doctors who must rely on it.
Until then, maybe the most provocative thought is this:

The cynicism of clinicians is not the problem in medicine. It might be the solution.

NOTES & ISSUES
in the real-life of dosimetry & clinical data

We invite the audience to suggest issues in dosimetry that

typically limit/hinder statistical analysis,

and/or points that deserve special attention.

NOTES & ISSUES
in the real-life of dosimetry & clinical data

Sample size - Very often, the number of patients undergoing dosimetry is limited by «availability», not established by statistical criteria, considering the possible interpatient variability, etc.

The AD data are often not regularly dispersed over the whole range of interest to determine effects, e.g., toxicity limits.

Trend vs. real correlation - often, a trend can also represent the most relevant result for clinical application, although a real statistical correlation is not derived; e.g., besides AD, many influencing factors can concur; further differentiation between tumor stage/grade, or patient status, etc., might be necessary and encouraged thanks to such first results.

The time interval for toxicity, etc. often, the data related to response and toxicity are not provided for an exact time interval of observation, thus patients with 2 months of response are mixed with 1 yr response, etc.
Detailed information is often omitted.

more...

NOTES & ISSUES
in the real-life of dosimetry & clinical data

The time interval for PFS, response, often, the data related to response and toxicity are not provided for a same time interval of observation, thus patients with 2 months of response are mixed with 1 yr response, etc. Detailed information is often omitted

Impact of lesion volume – e.g., the AD that allows efficacy might depend also on the lesion volume (effect of structure); multicollinearity?

The use percentages as independent variable - In some cases, especially in MRTs under development, the toxicity criteria are still to be clearly set; non rarely, the variation of clinical parameters vs. baseline can represent an alert, even if absolute values are still in «safe» ranges; thus, %variations are non rarely explored

The importance of REPORTING

more...

NOTES & ISSUES
in the real-life of dosimetry & clinical data

Other suggestions? Topic?