JOURNAL REHABILITÁCIA
--
LCRH
ISSN 0375 – 0922 www.journal.rehabilitacia.sk

Step by Step Manual Calculation of Descriptive Statistics, Confidence Intervals, Correlation Analysis, T-tests, F-test, Multivariate Analysis, Kaplan Meier Curves, and Log Rank Test in the Black Box Software Era for Transparent Evaluating Patient Health Status, Rehabilitation Outcomes, and Survival

[1]Daniel FURKA, [1]Saul DUEÑAS CASAS, [1]Andrej TÓTH, [1]Jerguš VENGRÍNI, [2]Ondrej LAURINEC, [3]Samuel HORVÁTH, [4]Anton GÚTH, [5]Karol HORNÁČEK, [1,6]Ivan FURKA, [1,7]*Samuel FURKA

[1]Comenius University, Faculty of Natural Sciences, Department of Physical and Theoretical Chemistry, Mlynská dolina, Ilkovičova 6, 84104-Bratislava, Slovakia.
[2]Department of Physical Medicine and Rehabilitation, University Hospital Bratislava, Hospital of Acad. L. Dérer, Limbová 5, 833 05 Bratislava, Slovakia.
[3]Landesklinikum Mödling, Landesgesundheitsagentur Niederösterreich, Restituta-Gasse 12, 2340 Mödling, Austria.
[4]Rehabilitation Institute Rehabilitácia - Professor Gúth Ambulatory Clinic, Na Barine 16, 841 03 Bratislava, Slovakia.
[5]Slovak Medical University in Bratislava, Faculty of Medicine, Clinic of Physical Medicine, Balneology and Medical Rehabilitation, Bratislava, Slovakia
[6]BEKO Engineering & Informatik GmbH, Stubenbastei 2, 1010 Wien, Austria.
[7]Department of Molecular and Biochemical Genetics, Institute of Medical Biology, Genetics and Clinical Genetics, Faculty of Medicine of Comenius University, University Hospital, Mickiewiczova 13, 813 69 Bratislava, Slovakia.

**Corresponding author: samuel.furka@uniba.sk*

SUMMARY/ABSTRACT

Starting point: Modern statistical models are essential for methodologically sound evaluation of patient health status changes and precise prognostication. Between 2003 and 2025, the scientific community has seen a rise in incorrect statistical application. Reliance on "Black Box" software, artificial intelligence, or automation without thorough understanding of key parameters often results in inaccuracies. Advanced statistical software streamlines calculations but does not replace model comprehension.
Group: We evaluated health status changes before and after treatment and rehabilitation in two distinct populations: wrist fracture patients and those with intramedullary osteosarcoma. Post-fracture muscular recovery demands precise diagnostics, particularly the standardised Janda test for reliable progress monitoring. Conversely, osteosarcoma rehabilitation prioritises verticalization training to optimise centre of gravity control and activate postural muscles. This approach maintains maximal sample heterogeneity, allowing monitoring of changes in both non-life-threatening and life-threatening diseases.
Methods: We propose a methodological approach incorporating a value matrix based on fuzzy logic at various significance levels to assess subjective, quasi-subjective, and objective health parameters. Consequently, we have detailed step-by-step manual calculations of descriptive statistics, confidence intervals, correlation analysis, t-tests, f-test, multivariate analysis, Kaplan-Meier curves and log-rank test, to ensure methodological clarity.
Results: Our study found that 65 out of 600 scientific papers employed inaccurate or not-optimized statistical models.
Conclusions: The aim of this work is to provide a methodological guide for the accurate application of advanced statistical models in scientific research, particularly for early-career physicians. Additionally, it aims to inform experienced professionals about potential ambiguities and procedural steps when utilising "Black Box" software.

KEYWORDS

Cancer, pain, musculoskeletal disorders, statistics & research methods, rehabilitation medicine.

DOI: 10.61983/lcrh.v63i2.167

1 INTRODUCTION

A wide range of parameters describing changes in a patient's health status requires the use of different statistical models in different situations. The correct categorisation of parameters and their evaluation is also the basis for scientific verification of hypotheses. Since mathematics also contains spicy nooks and crannies, it is crucial to understand statistical models and thus avoid bad or inaccurate interpretation of hard-earned results. Software tools such as Statcheck.io can be employed to verify the mathematical accuracy of calculated statistics[1]. A serious problem in the interpretation of statistics is the mathematically correct calculation using an improper set or inappropriate statistical model. From the researcher's perspective this may lead to the incorrect acceptance or rejection of the studied hypothesis and consequently to erroneous conclusions. From the perspective of a physician, current practice is based on evidence and medical science, requiring the active or passive acquisition of scientific information and verified research results. Therefore, understanding statistical methods as a fundamental part of medical research is essential, not only for evaluating the quality of final papers but also for correctly understanding the implications of research results. Unfortunately, in recent years, we have encountered very frequent inaccurate use of statistical models across chemical and medical fields. Sometimes these are minor errors that do not alter overall tendencies and trends, but other times they are critical errors. In accordance with general guidelines, we conducted a review of 600 scientific papers published in medical journals from 2003 to 2025. The articles were randomly selected and categorised based on their impact factor into five groups: a) IF: < 6, b) IF: 6 - 12, c) IF: 12 - 18, d) IF: 18 - 24, and e) IF: 24 - 30. From each group, we randomly selected 5 papers from three different journals for each year within the observed interval. The figure illustrates the frequency of the identified statistical issues (Fig. 1).

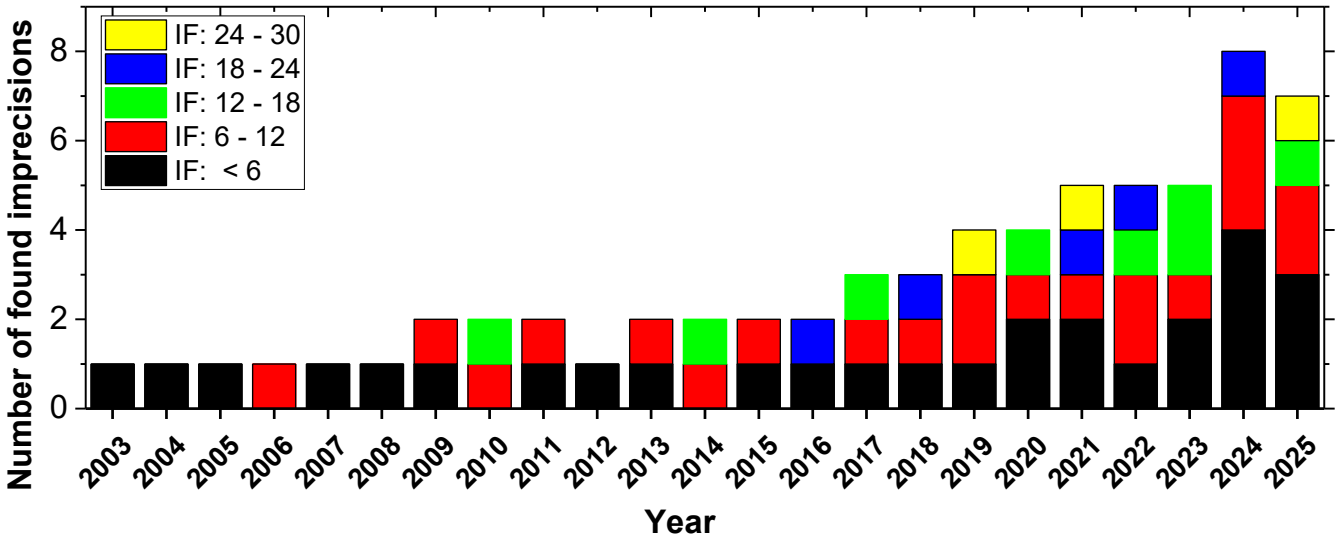


Figure 1 – Cumulative representation of the increase in the incidence of non-optimally used statistical models in medical-related scientific articles for different categories of impact factors (IF) in the period 2003-2025.

It was found that 65 out of 600 papers (10.8 %) contained incorrectly applied or sub-optimally selected statistical models, prompting us to write a study focused on the analysis of health status with methodological guidance and a step-by-step resolution of important statistical models. Interestingly, no numerical errors were found in any case, only suboptimal or even incorrect mathematical models. In instances where issues were identified in medical journals with an IF < 18, an optimal model would have led to a partial change in interpretation. For identified issues in journals with an IF > 18, optimisation would have resulted in changes to the presented deviations or critical values.

The foundation of statistical monitoring of changes in a patient's health status lies in the selection of relevant indicators, the method of information acquisition, and the choice of an appropriate mathematical model[2]. These three steps guide us through confusing and seemingly unrelated data to either reject or confirm our hypothesis. In practice, selecting relevant indicators is often simplified using software[3], published studies, or scoring tables[4].The score or indicator value classifies the patient into a disease subgroup, guiding treatment or rehabilitation according to protocol. Restoration of physiological joint play according to Lewit

in the region of the carpal bones represents a key benefit for ensuring the full range of motion and smooth recovery of hand function. Scoring tables and questionnaires are based on validated expert experiences. However, if no relevant scoring table exists for the monitored problem, we must identify the relevant parameters ourselves, focusing on those that most intensely depend on the disease's progression or regression[5].

For statistical evaluation, we can use basic descriptive statistics, including mean, mode, median, variability, and standard deviation[6] or, we can apply more advanced models like correlation analysis, covariance analysis, variance tests, and various T-tests[7–9]. T-tests compare the means of two groups, while variance tests assess variability within and between groups. More complex tests, such as ANOVA and MANOVA, enable the comparison of multiple groups. ANOVA (Analysis of Variance) compares the means of three or more groups, while MANOVA (Multivariate Analysis of Variance) extends ANOVA to compare multiple dependent variables across groups[10]. Modern statistical methods also allow us to track time-dependent phenomena, such as patient survival or implant longevity, using tools like the Kaplan-Meier curve or Log-Rank test. Employing appropriate statistical models for well-categorised patients provides clear insights into how treatments affect health outcomes[11]. Understanding treatment effects enhances our ability to refine strategies and predict results. Effective treatment strategies for specific patient categories can reduce costs, shorten treatment durations, prevent incorrect treatments, and improve prevention. A similar issue arises in diagnostic approaches. We need information on the diagnostic value of a method, including sensitivity, specificity, predictive values, and other parameters. The right diagnostic choice affects not only diagnosis accuracy but also cost, risks, and overall benefits. A good example is the screening process, which requires a method that is sensitive, widely available, and inexpensive. Effective screening should find the disease early enough to significantly influence patient survival while minimising false negatives and positives[12]. The complexity of intensive and time-consuming therapies requires long-term monitoring of various parameters, posing a challenge in evaluating the true improvement or deterioration of a patient's health. Therefore, we demonstrate the possibilities, advantages, and disadvantages of using individual statistical models to assess health status changes using real data from two vastly different patient groups: those with non-life-threatening conditions (wrist fractures) and those with life-threatening conditions (intramedullary osteosarcoma).

2 MATERIALS AND METHODS

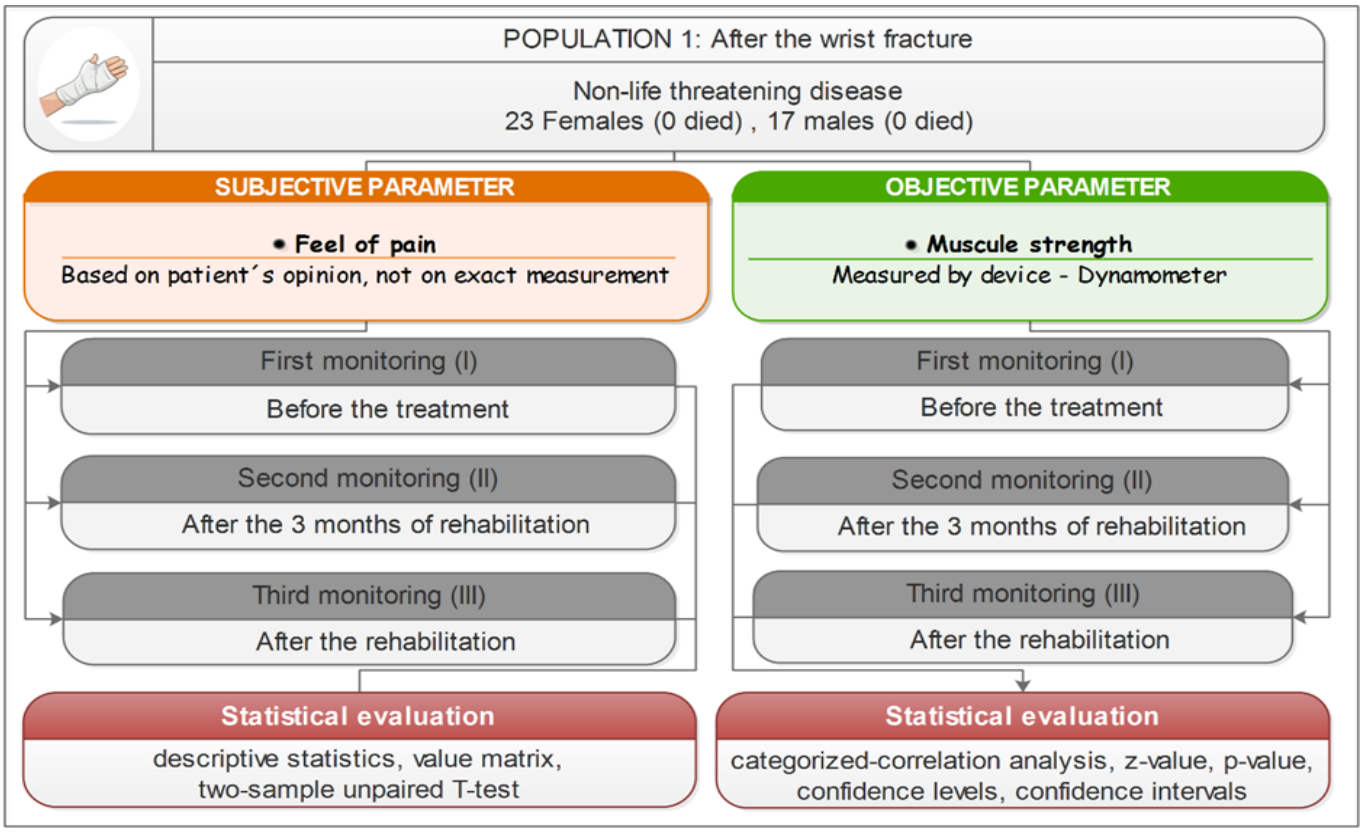
The aim of this study is to provide a practical overview of simple and advanced statistical models used in medical research, which increasingly relies on "Black Box" software. Understanding the manual calculations of models will give the doctor-researcher in the days of developing artificial intelligence an important control over his own confidence in the selection of monitored parameters, as well as in the subsequent selection of parameters, in the categorisation of data or during the process of statistical evaluation of the validity of hypotheses.

As objective parameters, we define parameters directly obtained by the measuring device or from laboratory results. These are completely independent of the patient and the doctor. Data recorded manually by a person, such as a doctor, are defined as quasi-objective parameters. These parameters may be influenced by error or the subjective experience of the physician, but not by the subjective view of the patient. In contrast, data obtained from the patient's statements are defined as subjective parameters. The collected data are evaluated using simple descriptive statistics, correlation analysis, variance tests, t-tests, Kaplan-Meier curves, and log-rank tests.

In a statistical investigation, it is important to distinguish whether the observed parameter is numerical or non-numerical[5]. For numerical parameters, the calculation involves a simple substitution into the mathematical model. For non-numerical parameters (such as verbal answers from patients), fuzzy logic is advisable. To implement this, we propose the "value matrix" method, which assigns discrete (integer) values to the various verbal responses within a question. Discrete values can be assigned according to the theoretical strength of the responses (e.g., no, I don't know, yes) or to responses without strength dependence (e.g., red, green, blue). Using the value matrix requires rounding the average response to a whole number for strength-independent answers and to two decimal places for strength-dependent answers. This method does not create a scoring table but instead, it provides real descriptive statistics from a single patient response. Value matrices are versatile and can be used anytime, but care must be taken to distinguish whether the

verbal responses are qualitative or quantitative. For this work, we studied three questions: (1) Do you feel pain in the limb when moving, (2) Do you feel pain in the limb at rest, and (3) Do you feel throbbing in the limb. Answers were assigned integer values according to fuzzy logic, with yes=1, maybe yes=2, I don't know=3, maybe no=4, and no=5. The transformed numerical data were then used for simple descriptive statistics.

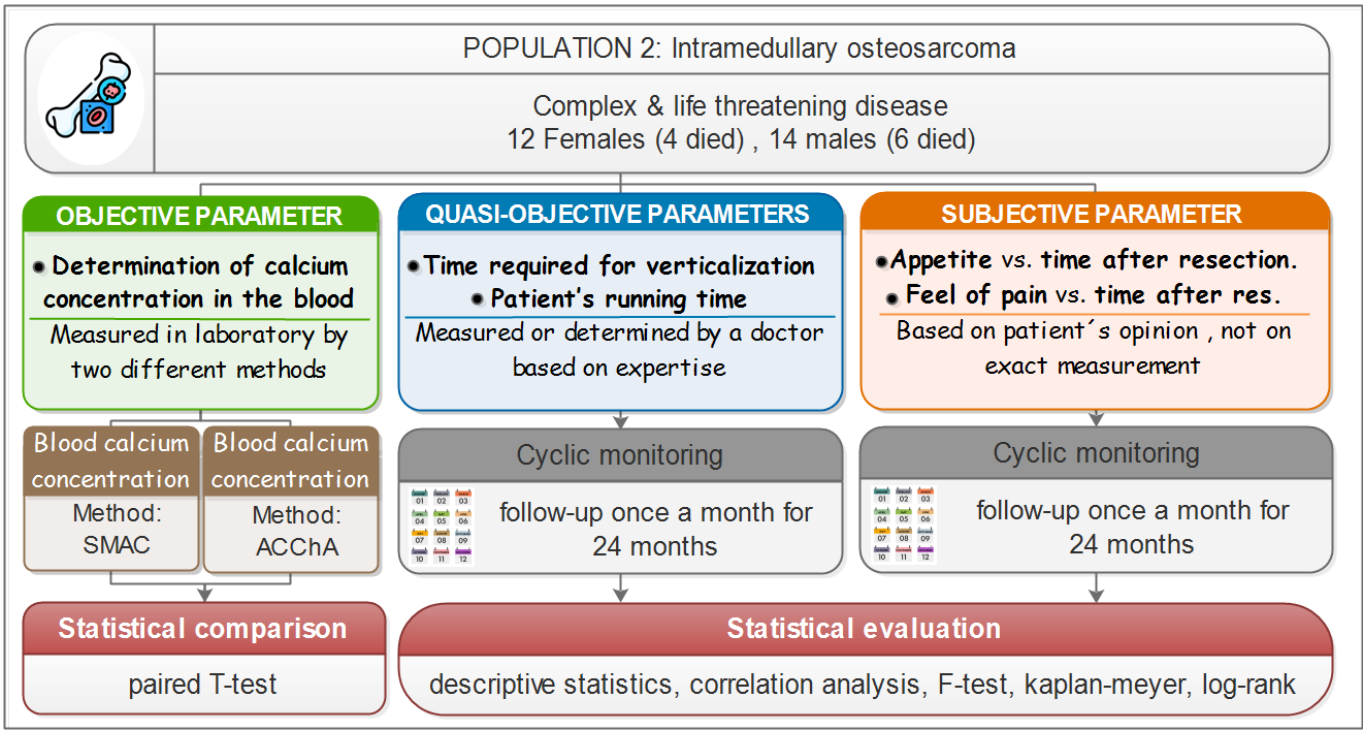
The first patient cohort consisted of 40 patients (23 females, 17 males) who had sustained a wrist fracture. This cohort represented patients with a non-life-threatening disease. All parameters were monitored: (I) before the treatment (II) after 3 months from the start of treatment during rehabilitation and (III) after the end of the treatment. A schematic representation of the structured map is provided in Scheme 1 (S. 1). The monitored parameters included pain intensity as a subjective measure and muscle strength assessed using a dynamometer as an objective measure. Early and targeted kinesiotherapy of the wrist after removal of immobilisation strongly supports tissue trophicity. Sensorimotor training using unstable surfaces together with specific occupational therapy aids produces highly positive results in the restoration of fine motor skills and grip dynamics.



S. 1 – Schematics showing a structured map of monitored parameters and statistical models used in the analysis of Patient Group 1 (After wrist fracture - non-life-threatening condition).

The second patient cohort, representing a complex and life-threatening disease, included 26 patients (12 females, 4 deceased; 14 males, 6 deceased) diagnosed with intramedullary osteosarcoma, whose parameters were monitored every month. In the case of this group, two patients were excluded because they withdrew from the study or failed to attend follow-up visits. A structured map is represented by a schematic (S. 2). The monitored parameters were the time required for patient mobilization to an upright position and the patient's running time on a 30-meter track two years after the beginning of the treatment as a quasi-objective parameter. The use of balneological procedures and hydrokinesiotherapy in thermal water markedly facilitates gait re-education due to water buoyancy, thereby maximising support for weakened lower limbs. The dosing of physical load in oncological patients strictly respects the current tolerance of the organism, thereby ensuring optimal energy balance and supporting overall vitality. As a subjective parameter, we chose the "feeling of pain" in the limb after physical exertion and "overall appetite" depending on the time after radical resection. As an objective parameter, blood calcium concentrations obtained using two

different laboratory methods were compared to determine whether the techniques produced statistically equivalent results.



S. 2 – Schematics showing a structured map of monitored parameters and statistical models used in the analysis of Patient Group 2 (Intramedullary osteosarcoma - life-threatening condition).

3 RESULTS AND DISCUSSION

3.1 SIMPLE DESCRIPTIVE STATISTICS AND VALUE-MATRIX

Simple descriptive statistics are most appropriate for small patient groups. Indicators such as the mean, mode, median, variability, and standard deviation help determine an individual's position within the set[13,14]. However, these statistics are insufficient for comparing large, heterogeneous groups of patients at a specific level of probability or significance. Simple statistics cannot definitively confirm or refute hypotheses regarding changes in a monitored parameter post-treatment, as they do not account for normal distribution. Nonetheless, they can effectively categorise patient groups. The terms "average" and "mean" are often used interchangeably. The average specifically refers to the arithmetic mean (sum of values divided by the number of measurements), whereas the mean can include other types, such as geometric and harmonic means. The K-means algorithm, often used in combination with artificial intelligence, is used for clustering and comparing large datasets[15]. Common descriptive statistics can be quickly calculated using a calculator or Microsoft Excel® (Tab. 1).

Table 1 – The most common quantities of simple descriptive statistics.

Quantity	What result do we get?	Syntax in MS. Excel®
Mode	The most frequent value of the measured parameter	=MODE(select_values)
Median	Determines the parameter value from which half of the data is greater than and half of the data is less than	=MEDIAN(select_values)
Arithmetic mean	The average value of a parameter calculated as the sum of the values divided by the number of observations	=AVERAGE(select_values)
Standard deviation (*SD*)	Precision, efficiency, or accuracy. SD describes how the values of the measured parameter differ from the average.	=STDEV.P(select_values)
Variability	A measure of variability, by multiplying by ^{2} we get the standard deviation	=VAR.P(select_values)

* *The quantity is given only for the needs of advanced statistics.*

When applying descriptive statistics to patient data, it is crucial to use multiple descriptive quantities for a comprehensive description, as relying on a single measure may be insufficient[16]. Second table displays the descriptive statistics for the "feeling of pain" parameter in patients after wrist fracture treatment and intramedullary osteosarcoma resection (Tab. 2). For instance, the most frequent response (mode) for wrist fracture treatment before the intervention is 1, indicating that most patients reported pain before treatment. If the mode and median are numerically similar, it suggests a normal distribution of the parameter, suitable for more complex analyses. Conversely, a significant difference between the mode and median indicates a non-normal distribution, affecting the choice and setup of advanced statistical models[16]. If the mode approximates the mean, the arithmetic mean is a suitable descriptor for the "feeling of pain." The standard deviation provides insight into the precision of patients' pain estimates relative to their group's average. Comparing descriptive statistics before, during, and after treatment shows a gradual decrease in the "feeling of pain" over six months. A comparison of patients revealed that those undergoing osteosarcoma resection experienced a longer recovery period compared to those undergoing wrist fracture rehabilitation.

Table 2 – Indicators of descriptive statistics of the subjective parameter "feeling of pain" on a sample of patients with different diagnoses before treatment, during rehabilitation and after the end of treatment.

	Patients with a wrist fracture			Patients with intramedullary osteosarcoma		
	Before treatment	**3 months after treatment begin**	**After treatment**	**Before treatment**	**3 months after treatment begin**	**After treatment**
Mode	1	2	4	1	1	4
Median	1	2	4	1	1	3
Average	1.4	2.3	3.75	2	1.3	3.1
SD	± 0.5	± 0.9	± 1.2	± 1.3	± 0.63	± 1.1

* 1=yes, 2=maybe yes, 3=don't know, 4=maybe no, 5=no

But what does the term recovery mean, and what significance does it hold in the case of a serious disease where some patients are cured, but a significant number die? Simple statistics cannot provide this answer in one step, thus, more complex models are necessary[17–19]. Interpreting simple statistics can be challenging and may lead to incorrect conclusions, particularly with complex therapies and confounding factors[17,18]. In such cases, numerical values might be misleading, requiring experience for accurate interpretation. As the patient number increases, the complexity and interpretability of simple statistics also grow. Advanced statistical models, based on normal population distribution, simplify interpretation by confirming or refuting hypotheses with a high level of probability. These models integrate multiple simple statistical indicators into a dimensionless number, representing a normal patient population[20]. They can confirm or refute hypotheses at various probability levels by incorporating a normal distribution of expected parameter values and considering small patient numbers with extreme parameter values, which may result from factors such as hypochondria, hypersensitivity, or reduced sensitivity. Advanced models produce a dimensionless number that combines multiple simple statistical indicators to represent a normal patient population. However, this approach sacrifices detailed numerical insights, such as average age or average muscular strength. Instead, we obtain information about the validity of our assumptions through specific "yes" or "no" outcomes. The foundation of complex statistics lies in confidence intervals, based on a distribution function, which define the maximum allowable upper and lower deviations of the measured parameter values[21].

3.2 SYMMETRIC AND ASYMMETRIC CONFIDENCE INTERVALS

The upper and lower deviation of measured parameters from the expected average value can be defined using either the standard deviation (SD) or the confidence interval (CI). These metrics allow us to distinguish between two sets of patients with identical modes, medians, and arithmetic means but differing in measurement accuracy or efficiency. Although both SD and CI indicate precision, they provide different interpretations. SD indicates the average deviation from the mean, where smaller values indicate more precise measurements or less variability between patient values. In contrast, CI describes the range within which the measured value is likely to fall with a certain probability, smaller values indicate a more precise estimate, reflecting higher reliability. The confidence level (CL) determines the probability of the CI and is typically set at 95% in medical statistics[22]. Lower confidence levels allow for more deviations in the

calculation. In medical applications, CI is often calculated as a distribution function of the mean (CI_{mean}), assuming a normal distribution of the monitored parameter, described by a Gaussian curve encompassing the most probable and least probable measured values (Eq. 1).

$$CI_{mean} = \bar{x} \pm \frac{z \cdot SD}{\sqrt{n}} \quad \text{(Eq. 1)}$$

Where $\bar{x}$ is the sample mean, n is the sample size, and z is the z-score corresponding to the desired confidence level. The z-score, a standardized value, indicates the number of standard deviations from the mean for specific confidence intervals. Common z-values and corresponding p-values are shown in table (Tab. 3). The p-value, often displayed in scientific publications, represents the probability of error, with the p-value multiplied by 100% indicating the error probability.

Table 3 – **Z**-values for calculation of a confidence intervals on a different confidence level.

Confidence level	80 %	90 %	95 %	99 %
***z*-value**	1.282	1.645	1.960	2.576
***p*-value (two-tailed)**	0.2	0.1	0.05	0.01

Confidence intervals are crucial in advanced statistical models to determine which measured values fit within the interpretation range. In this study CI_{mean} for muscular strength were determined by a dynamometer (Kern MAP 130K1) in a sample of 40 patients (23 females, 17 males). This method provides objective measurements, free from subjective biases. Calculated CI_{mean} for different confidence levels are presented in table (Tab. 4).

Table 4 – calculated values of the lower and upper limits of CI for two different confidence levels. Values represent hand muscular strength measured with a dynamometer.

Gender	**Necessary quantities**			**Confidence level 99 %** (z=2.576, p=0.01)		**Confidence level 95 %** (z=1.960, p=0.05)		**Confidence level 90 %** (z=1.645, p=0.10)	
	n	$\bar{x}$	SD	CI_{upper}	CI_{lower}	CI_{upper}	CI_{lower}	CI_{upper}	CI_{lower}
Males	17	346	42	372	320	366	326	362	329
Females	23	291	60	323	258	316	266	312	270

*Units = Newtons

The figure (Fig. 2) illustrates the difference in CI_{mean} and their reduction with decreasing CL. Red, green, and blue bars represent CIs for 99% (p = 0.01), 95% (p = 0.05), and 90% (p = 0.10) confidence levels, respectively. The height of the columns indicates the arithmetic mean of muscular strength for males and females.

In the figure 2, standard deviation represented as ± intervals around the mean, remains constant regardless of the confidence level (CL) and includes all potential probabilities and extreme errors. To accept a 1% error probability, we calculate the CI_{mean} at a 99% CL (Fig. 2). The resulting CI_{mean} is smaller than the original interval defined by the SD. As the CL decreases from 99% to 95% and finally to 90%, the CI_{mean} narrows, excluding more data points and admitting more errors. This shrinking CI_{mean} indicates which patients' data remain statistically reliable despite measurement errors. In our research, the CI_{mean} on the vertical axis defines the maximum and minimum acceptable muscular strength values. Each CI_{mean} implicitly includes the variability described by the SD, eliminating the need for separate error bars for SD. For accurate data interpretation, we must decide whether to show the dispersion of all measurements around the mean (SD) or define an interval of reliable values (CI), excluding certain error data among the patients.

A common mistake in professional papers is implicit assuming symmetrical CI_{mean} distributions around the mean. Symmetrical confidence intervals are typically obtained when the sampling distribution of the estimator is assumed to be symmetric, such as under normality assumptions, but other situations may require different distribution functions, resulting in asymmetrical CIs.

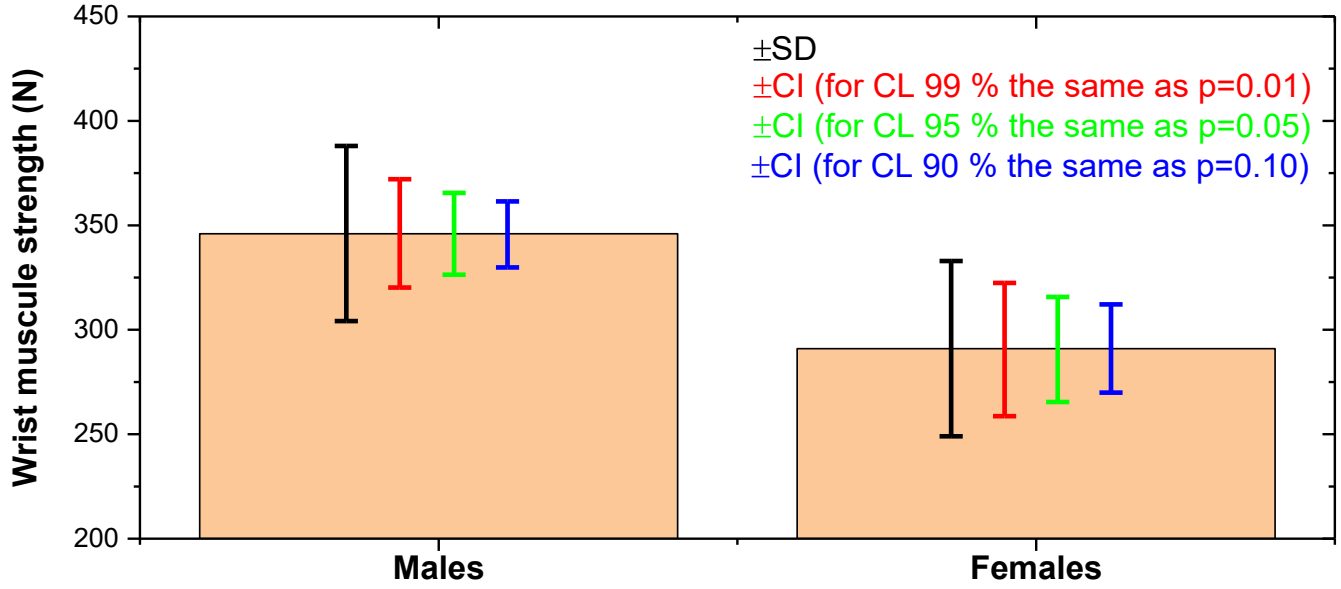


Figure 2 – Averages $\bar{x}$ of muscular strength among males and females before wrist rehabilitation (columns) together with calculated confidence intervals based on the distribution function of mean. Calculations were carried out for a three different confidence levels red) 99 %, green) 95 %, blue) 90 %. Calculated confidence intervals delimiting the confidence region of the data with different probabilities are compared with the black) standard deviations.

If the data do not follow a normal distribution, non-parametric confidence intervals should be used. One common method is bootstrapping (iterative method), which does not assume a specific known distribution and calculates the $CI_{bootstrap}$ by repeatedly resampling the dataset with replacement and computing the desired statistic. This calculation can be easily provided by statistical software. When dealing with small sample sizes, the t-distribution is used instead of the z-distribution. The confidence interval for the mean with a t-distribution CI_t is calculated using the equation 2.

$$CI_t = \bar{x} \pm t \cdot SD / \sqrt{n} \qquad \text{(Eq. 2)}$$

Where t is the t-score corresponding to the desired confidence level and degrees of freedom (f_{deg}). Degrees of freedom can be easily calculated as the sample size or number of samples n using the formula $n-1$. For proportions or percentages, the Wilson score interval provides a more accurate confidence interval, especially with small sample sizes or extreme proportions. While the Wilson confidence interval CI_{Wilson} provides better coverage than many other methods, its calculation is more complex, involving a quadratic equation and a more intricate solution process (Eq. 3), but is relevant also for small groups.

$$CI_{Wilson} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \qquad \text{(Eq. 3)}$$

Here, $\hat{p}$ is the sample proportion, n is number of samples, and z is the z-score corresponding to the desired confidence level. Interestingly, SPSS does not include a specific procedure for this calculation, but it allows it using commands. Variance confidence intervals, for variance and standard deviation require the chi-square χ^2 distribution and is defined as interval between the points (Eq. 4).

$$CI_{variance} = (lower, upper) = \left(\frac{(n-1)SD^2}{\chi^2_{lower}}, \frac{(n-1)SD^2}{\chi^2_{upper}}\right) \qquad \text{(Eq. 4)}$$

Where χ^2_{lower} and χ^2_{upper} are the chi-square critical values for the desired confidence level with $n-1$ degrees of freedom. In cases where data are skewed, transformations such as log transformation can normalize the data before calculating the confidence interval. Alternatively, the CI profile likelihood method can be used in statistical software. In biological and medical data evaluations, we recommend displaying

CIs at a specified confidence level rather than a generalized ± SD interval. This approach ensures a clear understanding of the range within which true parameter values are likely to fall, providing a more accurate and reliable basis for statistical interpretations.

Survival analysis often uses Greenwood confidence intervals CI_G. The Greenwood formula provides a measure of the precision of survival estimates in Kaplan-Meier survival analysis by estimating the variance of the survival probability at each time point. This variance is used to construct confidence intervals around the Kaplan-Meier survival estimates, indicating the range within which the true survival probability is likely to lie. The formula calculates the variance based on the number of patients at risk and the number of events (e.g., deaths) at each time point. Using the normal approximation, the confidence intervals are defined by equation (Eq. 5):

$$CI_G = S \pm Z\sqrt{Var_G} \quad \text{(Eq. 5)}$$

Where S is the Kaplan-Meier survival estimate at specific time, Var_G is the variance of the survival estimate at this time calculated using the generally known Greenwood formula and Z is the Z-value from the standard normal distribution corresponding to the desired confidence level. These intervals provide a statistical range likely containing the true survival probability, quantifying the uncertainty in the survival estimates.

3.3 CORRELATION ANALYSIS: EVALUATING TREATMENT EFFECTIVENESS AND HEALTH IMPROVEMENT

The methods of correlation analysis allow us to confirm two fundamental assumptions. First, correlation analysis helps determine whether a dependence exists between a time-dependent trend and a monitored parameter. Second, it provides information on whether the evolution of the parameter is influenced by a defined factor[23]. In monitoring health parameters related to disease progression, correlation analysis can assess whether a given treatment influences a patient's health (direction of effect) and the strength of that effect[24]. Additionally, this method allows us to observe the effects on various subgroups defined by age, gender, or specific comorbidities. Although the same mathematical tools are employed in both scenarios, the interpretations differ. Correlation analysis in medical research can be categorised into several types. Logarithmic and exponential correlations are used when the relationship between variables is nonlinear and can be more accurately described using a logarithmic scale. This includes methods where electromagnetic radiation or acoustic waves interact with body structures that exhibit high and frequency-dependent scattering. Another approach is a combination of linear and logarithmic correlation, in the research of enzymatic reactions and receptor responses, where the obtained graph is divided into two or more different functions, each correlated separately. Linearising non-linear data can be achieved through mathematical transformations or curve fitting, but correlation analysis should primarily be reserved for clearly linear dependencies.

Linear correlation is commonly used in ophthalmology, medical imaging, endocrinology, oncology, epidemiology, cardiology, genetics and genomics, neurology, and pharmacology. It is applied for optical and low-scattering spectroscopic methods, biomarker validation, drug efficacy studies, and ECG analysis, where the relationship between variables is represented by a straight line as a function of signal intensity after interaction with the sample. The strength of linear corelation is represented by the slope of the trendline indicates the effectiveness of a treatment or the progression of a biomarker. The linear trendline display function represents a linear fit to the data and is defined by the equation 6.

$$y = slope \cdot x + intercept \quad \text{(Eq. 6)}$$

Here, the $\pm\, slope$ describes the rate of increase or decrease of the dependence, while the intercept indicates the estimated value at $x = 0$. Generally, a larger slope signifies a stronger dependence. The accuracy of parameter monitoring can be assessed based on the distance of the graph points from the trend line, described by the correlation coefficient. The most used linear correlation coefficients for linearly dependent data are Pearson's correlation coefficient, or the R-square coefficient of determination (R^2). For linearly

dependent data, obtaining R^2 is straightforward using the trend line display in MS Excel (Fig. 3). In medical research, R^2 thresholds confirm or refute correlation. Values close to ±1 indicate excellent linear correlation due to treatment, whereas values near zero imply no significant effect on the monitored parameter.

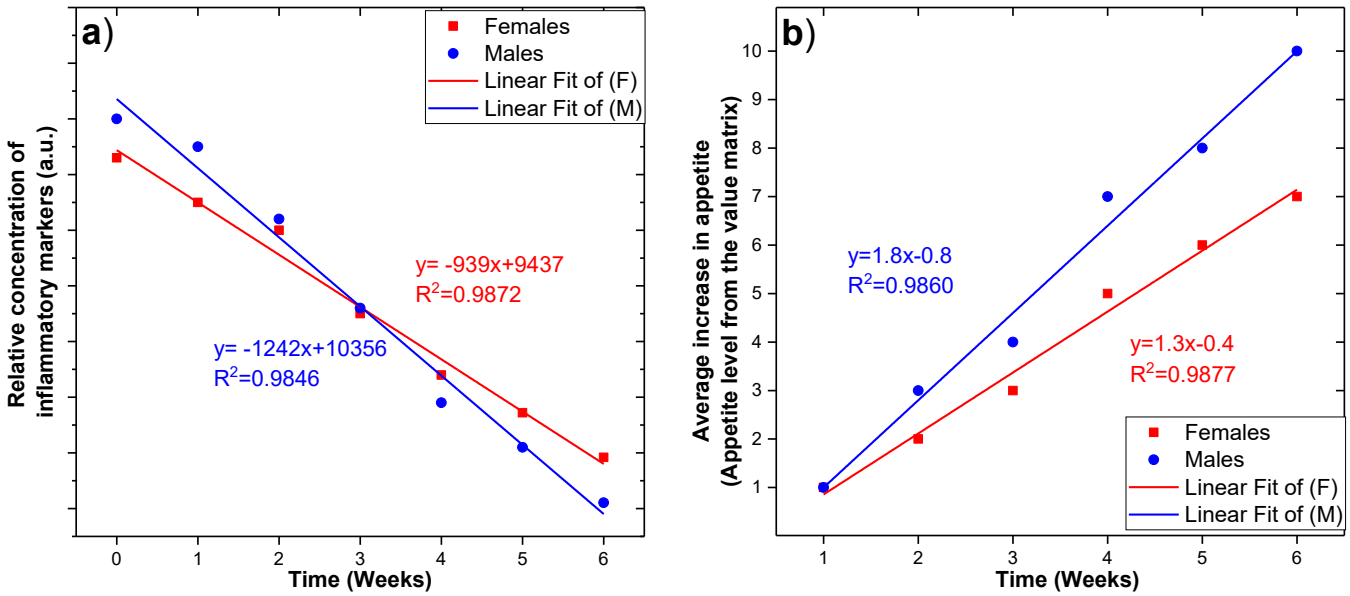


Figure 3 – Obtained correlations for patients after osteosarcoma resection a) average concentration of inflammatory markers in blood serum and b) increase in appetite as a function of time. In all cases, the correlation coefficient R^2 is close to 1, indicating excellent reliability.

For objectively obtained data (e.g., from a device or laboratory), correlation is confirmed when $R^2 \geq 0.95$. For quasi-objective data (estimated by expert), correlation is considered when $R^2 \geq 0.90$. For subjective data (patient-based answers), correlation should be confirmed for $R^2 \geq 0.80$. The threshold values decrease due to an increasing rate of errors and inaccuracies that obscure the true value of the parameter with decreasing objectivity of data acquisition[19,25]. It is important to note that correlation does not imply causation, because there may be confounding variables that affect the results. Therefore, more precise conclusions must be confirmed by further tests.

In the figure 3, we observe trend lines for the concentration of inflammatory markers and increased appetite over time after osteosarcoma resection for men and women. The decrease in inflammatory markers is 32 ± 2% more intense in men, despite similar R^2 values (Fig. 3a). Conversely, the increase in appetite, determined using fuzzy logic, is 27 ± 2% higher in men (Fig. 3b). The correlation coefficients for these trends may be similar but are unrelated to the trend line slopes.

Another situation is clustering. The graphical representation of hand muscular strength before and after wrist fracture treatment (Fig. 4) shows two clusters of points around the trend line and one cluster outside of it. These clusters can usually be matched with clinical data, indicating that muscular strength changed similarly in both subgroups during treatment. Points outside these clusters are considered errors or may represent extreme cases or unknown subsets with specific comorbidities. We can also estimate whether there is an improvement or deterioration of health. If there is no change, the measured values of muscular strength before and after treatment would be similar, represented by a trend line with a slope close to $\beta = 45°$ ($slope = tan\ \beta$). The more the slope differs from 45°, the more intense the health change is. For parameters that do not exhibit linear dependence, linear correlation cannot effectively evaluate health status improvement or deterioration.

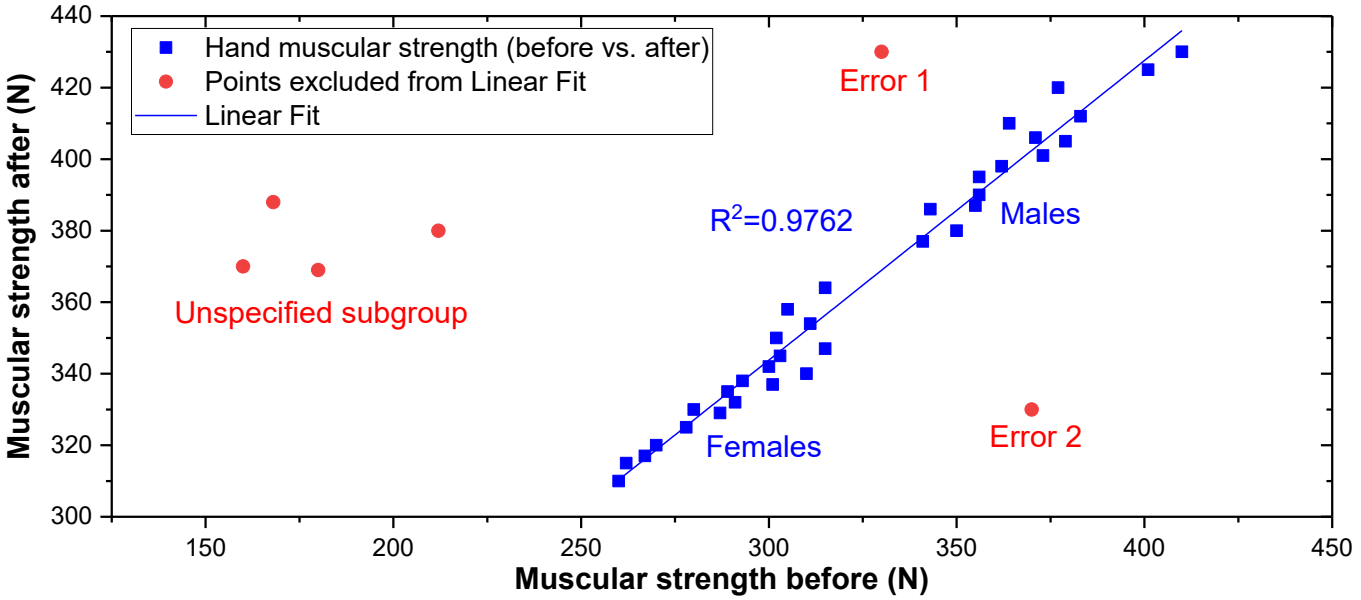


Figure 4 – Dependence of hand muscular strength before and after treatment of a wrist fracture. Two groups of points (blue) around the trend line are observable and belong to females and males. Excellent correlation $R^2 = 0.9762$ indicates improvement in muscular strength during treatment. Excluded points (red) represent obvious deviations from the trend line. It was not possible to categorize the unknown group according to the available data.

Figure 5 illustrates results where linear correlation is not applicable. In Fig. 5a, the absence of a systematic arrangement of points around a straight-line result in a low coefficient of determination, indicating no linear relationship. Conversely, Fig. 5b displays a reversed parabolic curve, suggesting a non-linear dependence despite an extremely low R^2. This implies a potential square root relationship rather than a linear one. When linear correlation is insufficient, more advanced methods are necessary. Partial correlation controls for additional variables, providing clearer insights into the relationship between two variables. Multiple correlation examines relationships among more than two variables, offering a broader understanding of complex interactions in medical data. Ensuring statistical significance is also crucial, typically evaluated using p-values to confirm that observed relationships are not due to random chance. By employing these advanced methods and rigorous statistical testing, a more accurate and comprehensive analysis of medical data can be achieved.

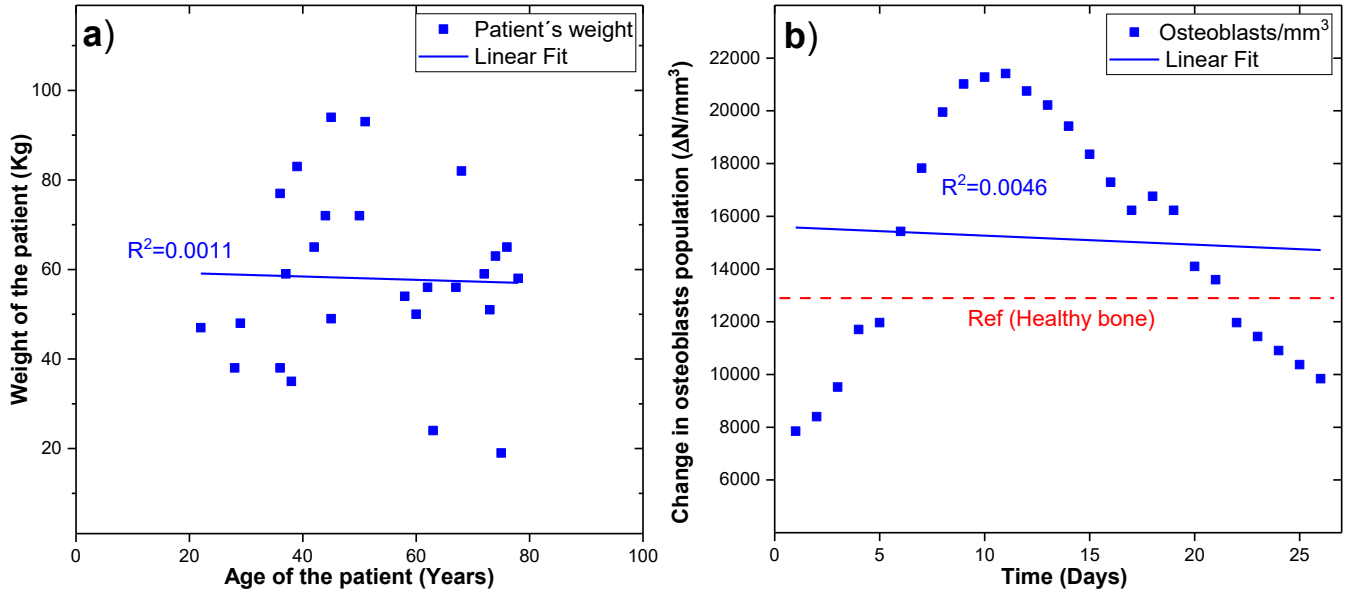


Figure 5 – The problems with a low correlation coefficient can be a) dependence of population age as a function of weight, where correlation coefficient is applicable and shows independence. On the contrary b) time dependence of cyclically applied magneto-therapy on osteoblast growth, which is non-linear but correct and therefore R^2 cannot be applied.

3.4 ADVANCED STATISTICAL ANALYSIS IN MEDICAL RESEARCH

Unlike simple descriptive statistics, which provide numerical information about samples or groups, advanced statistical models based on normal distribution confirm or reject a null hypothesis (H_0). Therefore, it is crucial to formulate the H_0 before initiating any calculations. A hypothesis can be any statement that will either be validated or unequivocally rejected. The boundary between the validity or rejection of the hypothesis is defined by the significance level (α), which is derived from the normal distribution of the general population[26].

The basic assumptions of a normal population include a heterogeneous population with normally distributed parameter values, encompassing both positive and negative extremes. According to the central limit theorem, the distribution of sample means approximates a normal distribution as the sample size increases, regardless of the population's distribution, given that the sample size is sufficiently large. In a normal distribution, α is a chosen threshold, typically set at conventional values like 0.1, 0.05, or 0.01. The smaller the significance level, the smaller the probability of observing a difference purely by chance. When using statistical tests based on normal distribution, it is crucial to validate the assumption of normality. This can be done using normality tests such as the Shapiro-Wilk test or by examining Q-Q plots. Correct selection of the significance level helps mitigate the impact of errors, outliers, and extreme values. In statistical research, we manually choose the threshold significance level. This value precisely determines the reference limit for rejecting the hypothesis, also known as the critical value. For measured objective parameters, a significance level of $\alpha = 0.05$ is recommended. For quasi-objective parameters obtained by a clinician, $\alpha = 0.1$ is suggested, while for subjective parameters defined by the patient and their perceptions, $\alpha = 0.2$ is recommended.

During hypothesis testing, it is essential to consider the possibility of errors. A Type I error occurs when a true H_0 is rejected (false positive), while a Type II error occurs when a false H_0 is not rejected (false negative). The α represents the probability of making a Type I error. The p-value indicates the probability of observing the test results under the H_0. A p-value lower than the chosen α suggests that the H_0 can be rejected, indicating statistical significance. The power of a statistical test is the probability of correctly rejecting a false null hypothesis. It depends on the significance level, sample size, and effect size. Higher power reduces the likelihood of a Type II error.

The selected critical value of the significance level α is then converted into a critical F-value (F_{crit}) and a critical T-value (T_{crit}) for a variance F-test and Student's T-test, respectively. The critical values of F_{crit} and T_{crit} for the same α have different values. These values can be easily obtained from relevant statistical tables or using Microsoft Excel® [27]. In addition to critical values, we need to calculate the descriptive F-value (F) and T-value (T) describing our datasets. These values result from specific combinations of simple descriptive statistics indicators. By comparing the critical values with our calculated T and F values, we determine whether our calculated values are lower than their critical values. If the T and F values are lower than their respective critical values, the datasets are statistically insignificantly different (subcritically matched). Conversely, if our values are greater than the critical values, we confirm that the sets are statistically significantly different. If these sets are significantly different, we can conclude that the treatment led to either an improvement or deterioration of the patient's health. When using the F-test and T-test, it is important to adhere to their respective assumptions. The F-test assumes homogeneity of variances, meaning the variances of the populations being compared should be equal. The T-test assumes equal variances and normality of the data. Note that Welch's T-test can be used if the assumption of equal variances is not met. In addition to statistical significance, it is important to consider the effect size, which measures the strength of the relationship between variables or the magnitude of the difference between groups. Common effect size measures include Cohen's d for the T-test and η-squared for the F-test. A statistically significant result does not always imply practical significance. While statistical significance indicates that an effect exists, practical significance assesses the real-world relevance or importance of that effect. We always consider the statistical set as a half-full and half-empty bottle, where confirming and refuting specific hypotheses can validate the same causality from either a negative or positive perspective.

3.4.1 F-TEST: EVALUATING OSTEOSARCOMA TREATMENT SUCCESS IN MALES AND FEMALES

To determine if the distribution of a parameter in two large groups of patients differs significantly, we employ variance testing, specifically the F-test. This method is suitable as we are not examining paired values, making it unnecessary for both groups to have equal sample sizes[28]. The F-test allows us to compare variances to identify differences in objective, quasi-objective, and subjective parameters. We conducted a variance test to analyse the verticalization ability of patients who underwent intramedullary osteosarcoma resection. The objective was to determine if there were significant differences in the ability to run between male and female patients two years post-surgery. Due to mortality and immobility, the sample sizes for males and females were unequal. Preliminary correlation analysis indicated gender differences in verticalization improvement, with males showing shorter and females showing longer durations. However, the impact of this difference on the ability to run was unclear. We hypothesized (H_0) that the recovery dynamics or efficiency between men and women in remission after intramedullary osteosarcoma treatment would differ. The calculation and results are presented in table (Tab. 5).

Table 5 – Statistics calculated from the running times of patients with different length of treatment.

Necessary quantities	Men (Running time/30m)	Women (Running time/30m)	Data calculation formula
Mean	$\bar{x}_2 = 9.46$	$\bar{x}_2 = 10.12$	*
Variance	$var_1 = 0.3301$	$var_2 = 0.4521$	*
Estimated variance	$var_{rel1} = 0.3405$	$var_{rel2} = 0.4689$	$= \frac{variance \cdot number\ of\ patients}{number\ of\ patients - 1}$
Number of patients	$n_1 = 8$	$n_2 = 6$	
Degrees of freedom	$f_{deg1} = 7$	$f_{deg2} = 5$	$= number\ of\ patients - 1$
$\boldsymbol{F}$ value	**1.3771**		$= \frac{var_{rel1}}{var_{rel2}}$
Critical value $\boldsymbol{F_{crit}}$	**3.3679**		• Calculation in MS. Excel® using =FINV(0.10;7;5) • Find data in statistic "F" table

* Equations and syntax for calculation in MS. Excel® are given in the chapter "simple statistics."

Given the quasi-objective nature of the stopwatch-based measurement, a significance level of $\alpha = 0.10$ was chosen. The calculated F was lower than the critical F_{crit}, indicating no significant difference between the groups (Tab. 5). Thus, despite differences in sample sizes and mean recovery times, the groups did not show significant differences at $\alpha = 0.10$. Both male and female patients demonstrated similar running performance regardless of the treatment duration. The F-test was performed by comparing group variances using their ratio. This ratio is compared to a critical value from the F-distribution table, determined by the chosen α and the degrees of freedom of each group. Additionally, it's important to verify the assumptions of the F-test, which include the independence of observations, normally distributed populations, and homogeneity of variances. If these assumptions are violated, the results of the F-test may not be reliable. The F-test is a crucial tool in ANOVA (Analysis of Variance) or MANOVA (Multivariate Analysis of Variance) methods, where it is used to compare variances and test for significant differences across multiple datasets. This makes the F-test versatile tool in analysing experimental data where comparison of variability is required.

3.4.2 T-TEST: EVALUATING DIFFERENCES BETWEEN PATIENT GROUPS

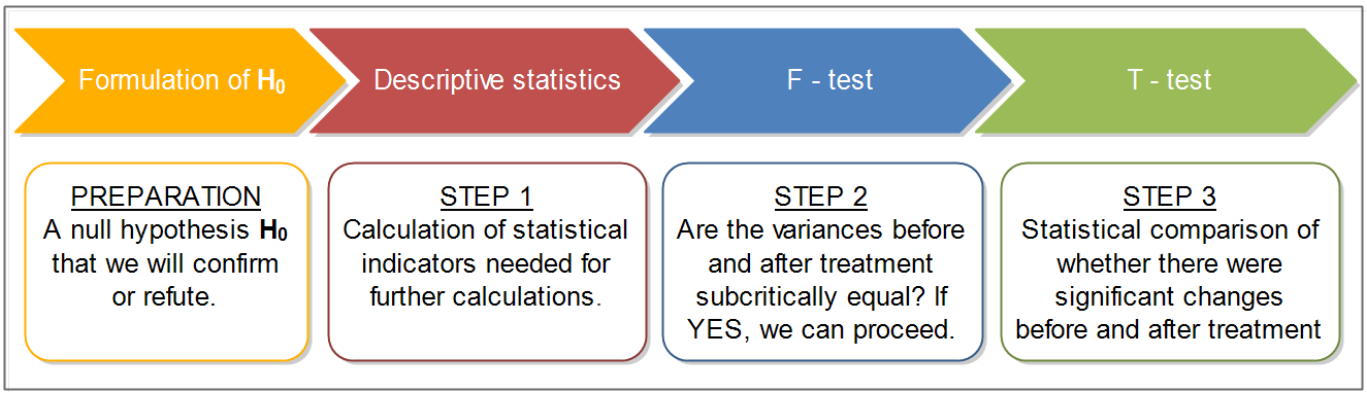
The T-test (Student's test) is used to compare the means of two data sets to determine if they are significantly different from each other. There are several approaches to the T-test, with the most common versions being the one-sample, two-sample (unpaired), and paired T-tests. The one-sample T-test compares the mean of a single group of patients to an expected reference value, determining if the mean of the sample significantly differs from the known or hypothesized population mean. The two-sample (unpaired) T-test compares the means of two independent groups of patients, assessing whether the means of two unrelated groups

differ significantly at a given α. The paired T-test compares two sets of values that are paired or matched in some way, such as pre- and post-treatment measurements for the same patients, requiring pairs of values strictly assigned to each patient to determine if there is a significant difference between the paired observations. Before choosing the appropriate type of T-test, it is essential to consider whether we are comparing the entire group against a reference value, against another independent group, or comparing paired values within a single group[29]. In contrast, T-test focuses on means, while the F-test focuses on variances. The F-test is often used in Analysis and Multivariate Analysis of Variance (ANOVA and MANOVA) to compare multiple groups. The paired T-test specifically handles paired data, a consideration not typically addressed by the basic F-test.

3.4.2.1 TWO-SAMPLE UNPAIRED T-TEST: ASSESSING CHANGES IN HEALTH STATUS

For comparing the similarity or difference between two groups of patients, a two-sample unpaired T-test is appropriate. This test can be used as a supplementary and verifying statistical model for correlation analysis. The two-sample unpaired T-test allows for the calculation of statistically significant changes during treatment. A key advantage of the T-test over correlation analysis is its purely mathematical foundation, which eliminates the need for graph evaluation. Additionally, unpaired T-test does not require pairs of values (before and after treatment) for each patient, as correlation analysis does. This flexibility allows for the inclusion of patients who did not return for a follow-up check-up or who died. It is not essential for the parameter values before and after treatment to belong to the same patient, as the test focuses on the overall group rather than individual patients. However, this test cannot be used to evaluate differences in the effectiveness of two different treatment methods, which would require monitoring paired data matched to individual patients.

For practical explanation we analyse studied dataset. Each of the 40 monitored patients with a wrist fracture underwent dynamometric measurements of strength in the affected hand. The first measurement was taken at the end of the third month of treatment (recorded in the first column in MS Excel®), and the second measurement was taken after the completion of treatment (recorded in the second column in MS Excel®). Since it is not possible to measure muscular strength immediately after the injury, the first measurement is considered "before treatment." The order of the patients is not important in this method, nor is the assignment of pairs of values to specific patients[30]. The calculation provides accurate results regardless of whether the two columns are of equal length, which can occur if some patients do not return for follow-up or due to their death. The independence from column lengths increases the robustness and applicability of the method. H_0 was: "After the treatment, the health status of the patients improved significantly." Calculation procedure is described in the subsequent steps that are visualised in the scheme (S. 3).



S. 3 – Schematic diagram mapping the procedure of obtaining the result when using the Two-sample unpaired T-test.

Step 1) We calculate basic descriptive statistics separately for the "before treatment" and "after treatment" columns in MS. Excel®. The calculated data are shown in the table (Tab. 6). Firstly, we will use an F-test, that verify the similarity of the variances of both columns (e.g., groups of patients or stages of treatment stages). If no statistically significant difference is observed, we can subsequently calculate the T-test corresponding to the formulated hypothesis H_0. It should be noted that there are specific cases where a comparison of F and F_{crit} can be used as a statistical model to evaluate the difference or similarity of two

groups. However, it answers another question, which is whether these two groups have the same or different variance of the monitored parameter. Thus, it can compare whether the group of females and males, the group of children and adults are similar. However, it does not tell us whether one group changed significantly compared to the reference group.

Table 6 – Statistics for dynamometrically measured hand muscular strength before and after wrist fracture treatment.

Necessary quantities	Calculated from col. I (3rd month after injury)	Calculated from col. II (After treatment)	Equation
Mean	$\bar{x}_1 = 314$	$\bar{x}_2 = 378$	*
Variance	$var_1 = 3574$	$var_2 = 2889$	*
Estimated variance	$var_{rel1} = 3666$	$var_{rel2} = 2964$	$= \frac{variance \cdot \text{number of patients}}{\text{number of patients} - 1}$
Number of patients	$n_1 = 40$	$n_2 = 40$	$=$ number of patients *in column*
Degrees of freedom	$f_{deg1} = 39$	$f_{deg2} = 39$	$=$ number of patients $- 1$
F value	**1.2371**		$= \frac{var_{rel1}}{var_{rel2}}$
Critical value F_{crit}	**1.7045**		• Calculation in MS. Excel® using =FINV(0.05,39;39) • Find data in statistic "F" table

* Equations and syntax for calculation in MS. Excel® are given in the chapter "simple statistics."

Step 2) Compare the calculated F value with the reference F_{crit} (Tab. 6). This step determines whether the variances of the two groups are statistically equal. If the variances of both columns are comparable. If $F > F_{crit}$, there is a statistically significant difference between variances. If $F < F_{crit}$, there is no statistically significant difference between variances and the columns are considered similar. For objective parameters measured by a dynamometer, $\alpha = 0.05$ is appropriate. The critical value F_{crit} can be obtained from the F-distribution table or calculated in MS Excel® using =FINV(0.05, 39, 39), where both 39 represents the degrees of freedom for each group. If F is lower than F_{crit}, proceed with the T-test. Calculate the T_{crit} value using MS Excel® with =TINV(0.05, 78), where 78 is the total degrees of freedom (sum of degrees of freedom for both columns). The result is $T_{crit} = 1.9908$.

Step 3) the equation for calculating the T value is quite complex, but we already know all the variables. Therefore, there is no need to worry, but with the help of a calculator and paper, we will quickly get the result. Term $|\bar{x}_2 - \bar{x}_1|$ denotes the positive difference between the averages, which we substitute into the formula (Eq. 7).

$$T = \frac{|\bar{x}_2 - \bar{x}_1|}{\sqrt{(n_1 \cdot {var_{rel1}}^2 + n_2 \cdot {var_{rel2}}^2)}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}} = 4.9366 \qquad \text{(Eq. 7)}$$

Where x_1 and x_2 are the means of each column, n_1 and n_2 are the number of patients in each group, and var_1 and var_2 are the variability estimates (Tab. 5). For our data, $T = 4.9366$ which is significantly higher than T_{crit} and the hypothesis H_0 is confirmed, indicating significant improvement in hand muscular strength at $\alpha = 0.05$. If the change is insignificant despite observable improvements, the monitored parameter or used α may be inappropriate. Increasing α, decreases accuracy and increases the risk of errors, so it is advisable not to exceed $\alpha = 0.30$. A lower α reduces the error rate and increases reliability.

3.4.2.2 PAIRED T-TEST: COMPARING EFFICIENCY OF RESULTS SIMILARITY OF TWO METHODOLOGIES

In postoperative care following the resection of intramedullary osteosarcoma or during the healing of fractures, monitoring calcium (Ca^{2+}) or calcitonin concentrations in the patient's blood is essential[31]. If paired data for each patient is unavailable, that patient must be excluded from the analysis. This requirement

for paired data is a fundamental difference from the two-sample unpaired T-test. The first option, which we did not study due to its similarity to previously addressed studies, includes monitoring calcium concentration at the beginning and end of postoperative treatment to obtain paired values for each patient. The second option, which we evaluated in this study, is to compare two methods of measuring blood Ca^{2+} concentration within a group of patients. In both scenarios, patients who missed the second check-up or died must be excluded due to the lack of paired values. The hypothesis H_0 was that both methods are equally representative. Blood Ca^{2+} concentration was measured using the Standard SMAC blood test (SMAC)[32] and the Advanced Clinical Chemistry Analyzer (ACChA)[33]. In MS Excel®, we created three columns: (I) measured calcium concentrations by SMAC, (II) measured concentrations by ACChA, and (III) the absolute differences between the two methods for each patient. If the subtraction results in a negative number, it is converted to a positive value. This column of positive differences is used for calculations as well as statistical parameters needed for evaluation of the paired T-test (Tab. 7). After calculating the differences of all pairs (rows), we average them to obtain the value $Dif_{average}$ shown in the table (Tab. 7).

Table 7 – Statistics calculated from blood Ca^{2+} concentrations determined by two different methods.

Necessary quantities	**Ca^{2+} concentration determined by SMAC**	**Ca^{2+} concentration determined by ACChA**	**Equation**
$\|\boldsymbol{Dif}\|$	Column (I)	Column (II)	$= \|value\ in\ (I) - value\ in (II)\|$ Repeat for each row separately
Average of positive differences	$Dif_{average} = 0.01$		$= AVERAGE(select_column\ III)$
Variance of differences $\boldsymbol{var_{dif}}$	$var_{dif} = 0.0003$		$= VAR.P(select_column\ III)$
$\boldsymbol{SD}$ of differences	$SD = 0.0183$		$= STDEV.P(select_column\ III)$
Number of patients	$n = 12$		
Degrees of freedom	$f_{deg} = 11$		$= number\ of\ patients - 1$
Critical value $\boldsymbol{T_{crit}}$	**2.2010**		• Calculation in MS. Excel® using =TINV(0.05;11) • Find data in statistic "T" table

Out of 26 patients diagnosed with intramedullary osteosarcoma, only 12 had blood tests performed by both methods and were included in the paired T-test calculation. Given the objective nature of calcium concentration measurements due to the accuracy of laboratory diagnostics, we chose a significance level of $\alpha = 0.05$. The final step is to calculate the T_{pair} value. The paired T-test relies solely on the column of differences (III), simplifying the calculation (Eq. 8).

$$T_{pair} = \frac{Dif_{average} \cdot \sqrt{n-1}}{SD} = 1.8166 \qquad \text{(Eq. 8)}$$

Substituting the descriptive statistics from Table 6, we calculated $T_{pair} = 1.8166$. This value is lower than $T_{crit} = 2.201$, indicating at a significance level of $\alpha = 0.05$, no significant difference was observed between the SMAC and ACChA methods in measuring blood calcium concentration in patients after resection of intramedullary osteosarcoma.

3.5 SUPPORTIVE THERAPEUTIC APPROACHES AND STATISTICAL INCLUSION OF THEIR EFFECTS

Additional supportive therapeutic approaches, including music therapy, hippotherapy, virtual reality assisted rehabilitation, mindfulness-based interventions and, in selected settings, art therapy or animal assisted therapy, are increasingly used as adjuncts to standard clinical care[34–38]. They should not be interpreted as substitutes for surgery, pharmacotherapy, physiotherapy or established rehabilitation protocols. Their expected contribution is mainly related to symptom burden, pain perception, anxiety, mood, cooperation, adherence, rehabilitation engagement, quality of life and selected functional outcomes. Therefore, their effects should be analyzed using the same methodological rigor as other subjective, quasi objective and objective parameters considered in this work.

Music therapy is evaluated statistically in the same general way as interventions in psychology, rehabilitation and clinical medicine. The primary aim is to determine whether a defined outcome changes after a specified number of therapeutic sessions, either within short term or long-term therapeutic goals, and whether this change differs from that observed in a control condition. Monitored outcomes may include eye contact, attention maintained for a defined time interval, participation in group activity, respect for rules, behavioural regulation, satisfaction with life, pain intensity assessed by a visual analogue scale, functional independence assessed by the Barthel Index, motor recovery assessed by Brunnstrom stages or spasticity assessed by the Modified Ashworth Scale. Music therapy is therefore most relevant for evaluating subjective and quasi objective outcomes related to pain perception, emotional regulation, attention, social interaction, motivation and cooperation with rehabilitation.

Hippotherapy is an adjunctive rehabilitation method based on the therapeutic use of horse movement. Rhythmic multidirectional motion may support postural control, trunk stability, balance, gait related function and sensorimotor integration, particularly in selected neurological and musculoskeletal rehabilitation settings. Its effects should be interpreted cautiously because the evidence is condition specific and protocol dependent. Therefore, hippotherapy should be evaluated using clearly defined functional outcomes, such as balance score, gait quality, sitting stability, trunk control, gross motor function and patient safety indicators.

Virtual reality assisted rehabilitation enables repeated task-oriented training with structured visual, auditory or sensorimotor feedback. This approach may improve patient engagement and support motor learning, balance training, gait rehabilitation, upper limb exercise and selected activities of daily living. Its effect can be assessed using objective parameters, such as range of motion, walking speed or task completion time, together with clinical scales and patient reported outcomes.

Mindfulness based interventions may improve coping with pain, anxiety, depressive symptoms, perceived stress, sleep quality and psychological adaptation to chronic disease. These effects should not be interpreted as direct evidence of accelerated tissue regeneration. Mindfulness based outcomes are most appropriately evaluated through subjective and quasi objective parameters, including anxiety score, pain score, perceived stress, sleep quality, treatment adherence and quality of life.

Art therapy and animal assisted therapy may be clinically useful in selected patients with reduced motivation, chronic stress, anxiety, social withdrawal or long-term rehabilitation needs. Their expected effects are usually related to emotional expression, social interaction, motivation, adherence, attention and perceived wellbeing rather than direct restoration of tissue structure. Their evidence base is more heterogeneous and outcome dependent, so their evaluation should remain limited to clearly defined behavioural, psychological or patient reported outcomes.

The statistical inclusion of supportive therapies should follow the scale type and study design. For continuous outcomes assessed in the same patients before and after therapy, the paired T-test may be used if the paired differences are approximately normally distributed. For each patient, the paired difference is defined as equation 9.

$$\boldsymbol{d_i = y_{2,i} - y_{1,i}} \quad \text{(Eq. 9)}$$

where $y_{2,i}$ is the post intervention value and $y_{1,i}$ is the pre intervention value. The mean paired change, standard deviation of paired differences, paired T-statistics are calculated using equations 10-12.

$$\boldsymbol{\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i} \quad \text{(Eq. 10)}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}} \quad \text{(Eq. 11)}$$

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} \quad \text{(Eq. 12)}$$

The number of degrees of freedom is $n - 1$. The 95 % confidence interval for the mean paired change is (Eq. 13).

$$\overline{\boldsymbol{d}} \pm \boldsymbol{t}_{\boldsymbol{1-\alpha/2,n-1}} \frac{\boldsymbol{s_d}}{\sqrt{\boldsymbol{n}}} \quad \text{(Eq. 13)}$$

The change should be reported together with its mean difference, confidence interval and standardized effect size. For paired measurements, the effect size based on difference scores is (Eq. 14).

$$\boldsymbol{d_z} = \frac{\overline{\boldsymbol{d}}}{\boldsymbol{s_d}} \quad \text{(Eq. 14)}$$

For small patient groups, the paired effect size may be corrected for small sample bias as (Eqs. 15-16).

$$\boldsymbol{g_z} = \boldsymbol{J d_z} \quad \text{(Eq. 15)}$$

$$J \approx 1 - \frac{3}{4(n-1)-1} \quad \text{(Eq. 16)}$$

where d_z is the standardised paired effect size and g_z is the small sample corrected paired effect size. If a therapeutic group is compared with a control group in a baseline and follow up design, ANCOVA or a mixed effects model is preferable to isolated independent T-tests, because these models directly test the intervention effect while accounting for baseline values and repeated measurements (Eqs. 17-18).

$$\boldsymbol{Y_{2,i}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 G_i} + \boldsymbol{\beta_2 Y_{1,i}} + \boldsymbol{\varepsilon_i} \quad \text{(Eq. 17)}$$

$$Y_{i,t} = \beta_0 + \beta_1 G_i + \beta_2 T_t + \beta_3 (G_i \times T_t) + u_i + \varepsilon_{i,t} \quad \text{(Eq. 18)}$$

For ordinal clinical scales or behaviorally coded categories, non-parametric or ordinal models are generally preferable. If the value matrix method is used as a study specific numerical transformation, its scoring rules should be explicitly justified and, where possible, verified by sensitivity analysis against an ordinal approach.

3.6 MULTIVARIATE ANALYSIS AS A TOOL OF ARTIFICIAL INTELLIGENCE

Multivariate analysis is a powerful tool in artificial intelligence, enabling the examination of multiple variables simultaneously to identify patterns, relationships, and trends[39–41]. Techniques such as Principal Component Analysis (PCA), Factor Analysis, Cluster Analysis, and Discriminant Analysis are widely used in AI to enhance data interpretation and decision-making. These methods help reduce data dimensionality, uncover latent structures, group similar observations, and classify data into predefined categories, thereby improving the accuracy and efficiency of AI models. However, each technique has its limitations, such as assumptions about data normality, sensitivity to outliers, and computational complexity, which must be carefully considered to ensure reliable results.

3.6.1 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis (PCA) reduces large datasets dimensionality by transforming correlated variables into a smaller set of uncorrelated principal components[42]. These components capture the maximum variance, simplifying complexity while retaining essential information. PCA is useful in medical imaging, such as MRI and CT scans, by reducing data complexity, and in genomics to analyse gene expression, identifying key genes influencing gene activity and aiding genetic disorder understanding. However, PCA has limitations, including potential loss of interpretability, the assumption that principal components with the highest variance are most important, sensitivity to the scale of data, and the assumption of linear relationships among variables, which may not perform well with non-linear data structures (Eq. 19).

$$PC_j = \sum_{i=1}^{p} w_{ij} \cdot X_i \quad \text{(Eq. 19)}$$

Where PC_j denotes the j-th principal component, constructed as a linear combination of the original observed variables X_i. The index i ranges from 1 to p, where p represents the total number of standardized clinical or biological variables included in the analysis. The coefficients w_{ij} are the component loadings, derived as elements of the eigenvectors from the covariance or correlation matrix $\mathbf{\Sigma}$, and they quantify the relative contribution of each X_i to the formation of PC_j. Components are extracted in descending order of explained variance, with PC_1 capturing the maximum possible variance, PC_2 capturing the maximum remaining variance under the constraint of orthogonality to PC_1, and so forth. These weights are derived by solving the eigenvalue decomposition $\sum w_j = \lambda_j \cdot w_j$, where λ_j represents the eigenvalue associated with the j-th component and corresponds to the amount of total variance explained.

3.6.2 FACTOR ANALYSIS

Factor Analysis identifies relationships between variables by modelling them as linear combinations of latent factors, uncovering underlying structures that explain observed correlations (Eq. 20). In clinical psychology, it can reveal core dimensions of mental health by identifying factors contributing to disorders from symptoms and questionnaire items, such as determining primary factors influencing anxiety and depression for targeted interventions[43]. However, factor analysis requires large sample sizes for reliable results, assumes linear relationships and accurate model representation, and can produce different solutions based on extraction and rotation methods. Additionally, the assumption of uncorrelated factors may not hold, and factor interpretation can be subjective and vary among researchers.

$$X_i = \sum_{j=1}^{m} \lambda_{ij} \cdot F_j + \varepsilon_i \quad \text{(Eq. 20)}$$

Where m indicates the number of underlying latent constructs ($m < p$), F_j represents the j-th unobserved common factor, assumed to follow a standard normal distribution. The parameters λ_{ij} denote factor loadings, reflecting the strength and direction of the relationship between observed variable i and latent factor j. The term ε_i captures unique variance attributable to measurement error or variable-specific influences not shared with other indicators. Factors are typically assumed to follow a multivariate normal distribution with zero mean and unit variance, and estimation commonly proceeds via maximum likelihood or principal axis factoring.

3.6.3 CLUSTER ANALYSIS

Cluster Analysis groups observations based on similarity without predefined categories. Common methods include k-means clustering[44], hierarchical clustering[45], and DBSCAN[46]. This technique is valuable for patient segmentation, such as grouping patients with similar symptoms or treatment responses to facilitate tailored medical interventions. For example, it can segment oncology patients based on their response to chemotherapy, enabling personalised treatment plans[46]. K-means clustering partitions data into k clusters, assigning each observation to the nearest mean[44] (Eq. 21). It is efficient for large datasets but requires specifying the number of clusters and is sensitive to initial centroids. Hierarchical clustering creates a dendrogram, representing nested groupings without needing a preset number of clusters, but can be computationally intensive. DBSCAN groups densely packed points and marks isolated points as outliers, finding arbitrarily shaped clusters without specifying the number of clusters but may struggle with varying densities and high-dimensional data. Cluster analysis has limitations, such as sensitivity to distance metrics and the number of clusters, varying results depending on the algorithm, and difficulty in interpreting and validating clusters, especially in high-dimensional data. It can be computationally intensive and requires proper data preprocessing and careful parameter selection. Additionally, cluster interpretability can be subjective, and the method may not perform well if the true data structure is not well-represented by the chosen model.

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|X_i - \mu_k\|^2 \quad \text{(Eq. 21)}$$

The objective function J represents the total within-cluster sum of squared Euclidean distances, which the algorithm iteratively minimizes. The parameter K specifies the predetermined number of patient subgroups. The notation C_k denotes the set of observations assigned to the k-th cluster during each iteration. The vector μ_k corresponds to the centroid, calculated as the arithmetic mean of all observations currently assigned to cluster k. The double vertical bars $\|\cdot\|$ indicate the Euclidean norm, measuring the geometric distance between an individual observation X_i and its corresponding cluster center.

3.6.4 DISCRIMINANT ANALYSIS

Discriminant Analysis classifies observations into predefined categories, commonly using Linear Discriminant Analysis[47,48] (LDA) and Quadratic Discriminant Analysis (QDA)[49,50]. In oncology, it can classify tumours as benign or malignant based on multiple diagnostic features, aiding in early and accurate diagnosis. For example, LDA helps pathologists categorise tumour samples, based on RNA sequencing data, improving cancer diagnosis and guiding treatment decisions. However, discriminant analysis assumes normality for predictors and equal covariance matrices in LDA and is sensitive to outliers. It may perform poorly with non-linear relationships or small sample sizes and can be less effective if the classes are not well-separated, leading to potential misclassification (Eqs. 22-23).

$$y_i = \sum_{j=1}^{p} w_j \cdot x_{ij} \quad \text{(Eq. 22)}$$

$$w = {S_w}^{-1} \cdot (\bar{x}_1 - \bar{x}_2) \quad \text{(Eq. 23)}$$

Where y_i denotes the discriminant score for observation i, computed as a linear combination of the original observed variables x_{ij}. The index j ranges from 1 to p, representing the total number of predictor variables included in the analysis. The coefficients w_j are the discriminant weights. In equation 13, w is the vector of discriminant weights (w_1, w_2, ..., w_p). The matrix S_w is the pooled within-class covariance matrix, which measures how much the predictor variables vary within each group. The term ${S_w}^{-1}$ is its inverse. The vectors $\bar{x}_1$ and $\bar{x}_2$ are the mean vectors of groups 1 and 2, respectively. The difference $(\bar{x}_1 - \bar{x}_2)$ represents the separation between groups for each predictor. Multiplying ${S_w}^{-1}$ with $(\bar{x}_1 - \bar{x}_2)$ yields weights that maximize between-group separation while accounting for correlations among predictors. Observations are classified into the group yielding the highest discriminant score.

3.6.5 MULTIVARIATE REGRESSION

Multivariate Regression extends multiple regression by modelling multiple dependent variables simultaneously[51,52]. This method reveals how independent variables affect several outcomes. In cardiology, multivariate regression can analyse the impact of lifestyle factors on multiple cardiovascular outcomes, such as blood pressure, cholesterol levels, and heart rate. This comprehensive analysis helps in understanding how various factors contribute to heart health, guiding both prevention and treatment strategies. However, multivariate regression has limitations, such as the assumption of multivariate normality and linearity, the need for large sample sizes to ensure reliable estimates, and potential multicollinearity among independent variables, which can complicate the interpretation of results. Additionally, the model's complexity increases with the number of dependent variables, which may lead to overfitting if not properly managed. Furthermore, the interpretation of the coefficients can be challenging due to the interdependencies among the dependent variables (Eq. 24).

$$Y_j = \beta_j + \sum_{k=1}^{p} \beta_{jk} \cdot X_k + \varepsilon_j \quad \text{(Eq. 24)}$$

Where Y_j denotes the j dependent variable, with q representing the total number of outcomes modeled simultaneously. The index k ranges from 1 to p, representing the number of independent variables. The parameter β_j is the intercept for outcome j, while β_{jk} represents the regression coefficient quantifying the change in Y_j associated with a one-unit change in predictor X_k, holding other predictors constant. The term ε_j captures the residual error for outcome j. Unlike univariate regression, this approach accounts for correlations among the q dependent variables, providing more efficient estimates and enabling joint hypothesis testing across multiple outcomes.

3.6.6 CANONICAL CORRELATION ANALYSIS (CCA)

Canonical Correlation Analysis (CCA) examines relationships between two sets of variables, identifying linear combinations that are maximally correlated[53]. In genomics[54], CCA explores links between genetic markers and phenotypic traits, uncovering significant genetic influences on physical characteristics and disease mechanisms. For example, CCA identifies genetic markers most strongly associated with specific diseases, aiding in therapeutic target discovery. However, CCA assumes linear relationships, is sensitive to outliers, requires large sample sizes for reliable results, and can be computationally intensive and complex to interpret. Proper data preprocessing and normalisation are crucial for meaningful results (Eqs. 25-26).

$$\begin{cases} U = \sum_{i=1}^{p} a_i \cdot X_i \\ V = \sum_{j=1}^{q} b_j \cdot Y_j \end{cases} \quad \text{(Eq. 25)}$$

$$r_c = correlation(U, V) = \frac{Cov(U,V)}{\sqrt{Var(U) \cdot Var(V)}} \quad \text{(Eq. 26)}$$

Where U and V denote the first pair of canonical variates (linear combinations). The set $\{X_1, \ldots, X_p\}$ represents the first variable block, while $\{Y_1, \ldots, Y_q\}$ represents the second. The coefficients a_i and b_j are canonical weights derived to maximize the correlation r_c between U and V. Subsequent pairs of canonical variates ($U_2, V_2, \ldots$) are extracted under the constraint of being uncorrelated with previous pairs. The maximum number of canonical correlations equals $\min(p, q)$. Statistical significance is typically assessed using Wilks' Lambda or Bartlett's chi-square test.

3.6.7 MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Multivariate Analysis of Variance (MANOVA) extends ANOVA by evaluating multiple dependent variables simultaneously, useful in experimental studies[55,56]. In clinical trials, MANOVA assesses a drug's impact on various health outcomes like blood sugar, weight, and cholesterol, providing a comprehensive view of treatment efficacy. However, MANOVA assumes multivariate normality, homogeneity of variance-covariance matrices, and requires large sample sizes. It is sensitive to outliers and complex to interpret with non-linear relationships. Proper data preprocessing and model assumption consideration are essential for reliable results (Eq. 27).

$$Y_i = \mu + \tau_g + \epsilon_i \quad \text{(Eq. 27)}$$

Where Y_i denotes the vector of dependent variables for observation i, with q representing the total number of dependent variables modeled simultaneously. The vector μ represents the overall mean vector across all groups. The parameter τ_g represents the effect of group g on the dependent variables. The term ϵ_i is the random error vector for observation i, assumed to follow a multivariate normal distribution with mean zero and covariance matrix Σ. MANOVA tests the null hypothesis that all group effects are zero using Wilks' Lambda statistic (Eq. 28).

$$\Lambda = \frac{|\boldsymbol{E}|}{|H+\boldsymbol{E}|} \quad \text{(Eq. 28)}$$

Where E is the error sum of squares and cross-products matrix, and H is the hypothesis sum of squares and cross-products matrix.

3.7 MONITORING OF A SURVIVAL

3.7.1 KAPLAN-MEIER CURVE – MONITORING THE SURVIVAL OF PATIENTS WITH IN-TRAMEDULLARY OSTEOSARCOMA

The Kaplan-Meier (KM) curve is a widely used statistical method in survival analysis, useful for handling censored data[57,58]. Censoring (C) occurs when patients withdraw from the study for reasons other than death or when the time of death is unknown. Censoring does not compromise the validity of the result[59]. This method estimates the probability of survival over time and can compare survival rates across different treatments or patient groups with varying comorbidities. The assumptions of the Kaplan-Meier method

include independent censoring, where the censoring mechanism must be independent of the survival prospects, and consistent entry and exit criteria, ensuring well-defined study parameters[60].

In our study, KM curve was used to monitor the survival of patients following resection of intramedullary osteosarcoma (Tab. 8, Fig. 6). Out of 26 patients, two were censored. Table 8 details the survival data, including time intervals for monitoring (W), the number of deaths within each interval (E), the number of live patients at the start of the interval (n), the cumulative survival rate for uncensored patients (S_{un}), the survival rate adjusted for censored patients (S_{cen}), and the iterative survival rate calculation for censored data (S_{it}).

Table 8 – Calculation of the Kaplan-Meier curve for patients with intramedullary osteosarcoma.

			UNCENSORED	CENSORED		
W (Month)	E	$n = n_{prev} - E_{prev}$	$S_{un} = \frac{n_{next}}{n_{all}}$	C	$S_{cen} = \frac{n-E}{n}$	$S_{it} = S_{cen} \times S_{it(prev)}$
0 - 4	0	n_{all} =26	26/26 = 1.00 = **100 %**	0	1.00	**100 %**
4 - 8	2	n = 26−0 = 26	24/26 = 0.92 = **92 %**	0	0.92	× 1.00 = 0.92 = **92 %**
8 - 12	4	n = 26−2 = 24	20/26 = 0.77 = **77 %**	1	0.83	× 0.92 = 0.76 = **76 %**
12 - 16	5	n = 24−4 = 20	15/26 = 0.58 = **58 %**	0	0.75	× 0.76 = 0.57 = **57 %**
16 - 20	0	n = 20−5 = 15	15/26 = 0.58 = **58 %**	1	1.00	× 0.57 = 0.57 = **57 %**
20 - 24	3	n = 15−0 = 15	12/26 = 0.46 = **46 %**	0	0.80	× 0.57 = 0.46 = **46 %**
24	1	n = 15−3 = 12	11/26 =0.62 = **42 %**	0	0.92	× 0.46 = 0.42 = **42 %**

The calculation involves determining the number of patients at risk at each interval, the proportion surviving, and then adjusting these proportions iteratively to account for censored data. The inclusion of censored patients in the analysis is crucial for a precise and reliable interpretation of survival rates. The survival rate at each interval is derived by multiplying the survival rate from the previous interval by the current rate for both uncensored and censored data. Figure 6 shows the Kaplan-Meier curve, illustrating the survival probabilities over time (Fig. 6). The curve shows minimal differences between uncensored and censored data, with both groups displaying a 42 % survival rate at 24 months. The median survival time (M_{surv}), identified as the time at which the survival probability intersects 50 %, is 20 months for both groups[59]. This value aligns with existing literature. The slight differences observed in survival rates between uncensored and censored data emphasize the need for this iterative approach (Tab. 8). The KM curve and median survival time provide insights into patient prognosis and the effectiveness of treatment interventions.

The Kaplan-Meier method handles ties (multiple events at the same time point) using statistical methods such as the Exact method, Breslow method, and Efron method. The Exact method assumes events occur in an infinitesimally small-time interval in a random order, making it computationally intensive due to the enumeration of all possible event orderings. The Breslow method, or generalized Wilcoxon method, simplifies calculations by equally distributing tied events across the risk set, averaging survival probabilities. The Efron method balances computational efficiency and accuracy by incrementally adjusting the risk set for each tied event, fractionally allocating the events for a more precise estimate. Confidence intervals for survival estimates, calculated using the Greenwood formula, provide precision understanding (Eq. 5). Right-censoring, the most common form, occurs when the study ends before the event for some subjects, or they are lost to follow-up. Left-censoring occurs when the event happens before the observation period starts, and interval-censoring occurs when the event is known to occur within a specific time interval. While the Kaplan-Meier method is excellent for descriptive purposes, the Cox proportional hazards model is often used for multivariate survival analysis, allowing the inclusion of covariates to understand their impact on survival[61]. Additionally, the log-rank test is commonly used to compare survival distributions between groups, assessing the significance of differences between KM curves.

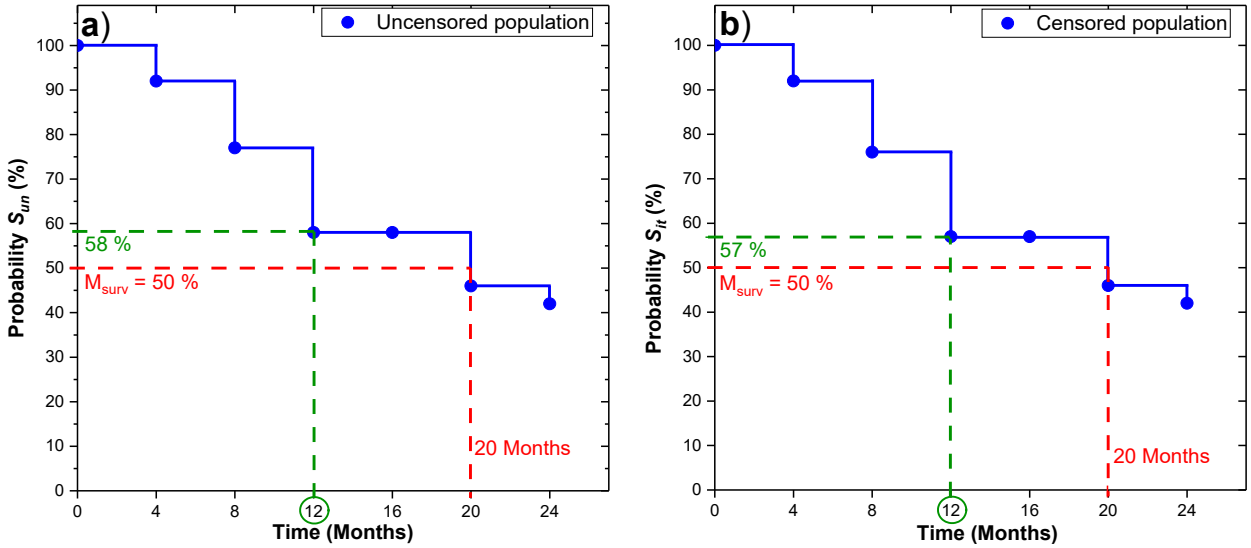


Fig. 6 – Kaplan – Meier curves for the same population of patients treated for intramedullary osteosarcoma where (a) represents uncensored group of patients, and (b) represents the same group of patients after censoring. The green intersection shows survival rate at 12 months. Median survival (red) corresponds to 50 % probability and takes the value of 20 months for both censored and uncensored patients.

3.7.2 LOG RANK TEST – COMPARISON OF TIME-DEPENDENT PHENOMENA OR THE SURVIVAL RATES OF SEVERAL GROUPS OF PATIENTS

To statistically evaluate and compare the survival rates of two populations, the Log Rank test (LR) is recommended for determining the difference or similarity in survival distribution functions over time[62]. This test complements the KM method by statistically confirming or rejecting the similarity of survival distributions. The LR test can compare other time-dependent events, such as the duration of treatment, rehabilitation periods, or the time-dependent lifetime effect of materials used in medical devices. A common application is comparing survival distributions between groups of patients with or without certain conditions or treatments. The LR test assumes that the hazard ratios between groups are constant over time (proportional hazards assumption). Violations of this assumption can affect the validity of the test results. LR also assumes that censoring is random and not related to the likelihood of an event occurring. Non-random censoring can introduce bias[63].

In our study, we aim to verify the hypothesis H_0: "Censoring less than 10 % of patients does not lead to a significant difference in survival time profiles." To test this hypothesis, we compare two groups of patients after osteosarcoma resection: $group\ 1$ without censoring and $group\ 2$ with censoring. Firstly, we create the table to track the number of surviving patients in each group over specified time frames (Tab. 9).

Table 9 – Preparatory table for Log Rank statistics, showing changes in the number of patients during the monitored period.

			Group 1 - UNCENSORED	Group 2 - CENSORED		
W (Month)	E_1	C_1	$n_1 = n_{prev} - E_{prev}$	E_2	C_2	$n_2 = n_{prev} - E_{prev}$
0 - 4	0	*0*	n_{all} = 26	0	*0*	n_{all} = 26
4 - 8	2	*0*	n_1 = 26−0−*0* = 26	2	*0*	n_2 = 26−0−*0*=26
8 - 12	4	*0*	n_1 = 26−2−*0* = 24	4	*1*	n_2 = 26−2−*0*=24
12 - 16	5	*0*	n_1 = 24−4 −*0* = 20	5	*0*	n_2 =24−4−*1*=19
16 - 20	0	*0*	n_1 = 20−5 −*0* = 15	0	*1*	n_2 =19−5−*0*=14
20 - 24	3	*0*	n_1 = 15−0−*0* = 15	3	*0*	n_2 =14−0−*1*=13
24	1	*0*	n_1 = 15−3−*0* = 12	1	*0*	n_2 =13−3−*0*=10

The first column, W, lists the time frames in months, starting with 0-4 months and continuing in four-month intervals. The number of rows corresponds to these intervals over a 24-month period. The second column, E_1, represents the number of patients who died in each interval. The third column, C_1, lists the censored patients. The last column, n_1, indicates the total number of living patients at the start of each

interval, including all patients present on the first day without removing those who died during that period. Patients who died in each interval are removed from the risk set in subsequent intervals. Table 9 illustrates these calculations. Similarly, columns E_2, C_2, and n_2 for $group\ 2$ account for both deceased and censored patients. The n_1 and n_2 values are determined as in the Kaplan-Meier method but adjusted by subtracting censored patients from the previous time frame.

Next, we construct a table for direct calculation of the Log Rank statistics (Tab. 10) using the data from the preparatory table 9. Table 10 includes Exp_1 and Exp_2, the expected number of deaths for each group, calculated using the equations (Eq. 29a, b).

$$Exp_1 = \left(\frac{n_1}{n_1+n_2}\right) \cdot (E_1 + E_2) \quad \text{or} \quad Exp_2 = \left(\frac{n_2}{n_1+n_2}\right) \cdot (E_1 + E_2) \qquad \text{(Eq. 29a, b)}$$

An important value is the Log Rank variance (var_{LR}), defined by equation (Eq. 30). Log Rank variance determines the degree of variability and depends on the deviation from the population mean. It is crucial to calculate the variance separately for each row of the table, as it is not a linear phenomenon.

$$var_{LR} = \frac{n_i \cdot n_2 \cdot (E_1 + E_2) \cdot (n_1 + n_2 - E_1 - E_2)}{(n_1 + n_2) \cdot (n_1 + n_2) \cdot (n_1 + n_2 - 1)} \qquad \text{(Eq. 30)}$$

Differences between observed and expected deaths for $group\ 1$ and $group\ 2$ are calculated as $Dif_1 = E_1 - Exp_1$ and $Dif_2 = E_2 - Exp_2$ respectively (Tab. 10). It is not necessary to calculate both values since they give us the same numerical value for the same row, differing only in the $\pm$ sign.

Table 10 – Calculated expected number of deaths (Exp_1, Exp_2), Log Rank variance values (var_{LR}) and differences (Dif_1, Dif_2) for all monitored time frames W.

	Expected deaths for each group		**Variance between groups**	**Observed – expected deaths**	
W ***(Month)***	$\boldsymbol{Exp_1}$ **(**For group 1)	$\boldsymbol{Exp_2}$ **(**For group 2)	$\boldsymbol{var_{LR}}$ (Common to both groups)	$\boldsymbol{Dif_1}$ **(**For group 1)	$\boldsymbol{Dif_2}$ **(**For group 2)
0 - 4	0	0	0	0	0
4 - 8	2	2	0.9412	0	0
8 - 12	4	4	1.7021	0	0
12 - 16	5.1282	4.8718	1.9066	− 0.1282	0.1282
16 - 20	0	0	0	0	0
20 - 24	3.2142	2.7857	1.2160	− 0.2143	0.2143
24	1.0909	0.9090	0.4723	− 0.0909	0.0909

After calculating the individual values for each time frame, we summarise them to obtain the total variance var_{LR_SUM} and total differences Dif_{1_SUM} or Dif_{2_SUM} as shown in Table 11.

Table 11 – Calculated Log Rank variables for the comparison of survival distribution functions of censored and uncensored patients after resection of intramedullary osteosarcoma.

Description	**Equation**	**Result**
The sum of the calculated variances $\boldsymbol{var_{LR}}$	$\boldsymbol{var_{LR_SUM}} = \sum_{i=1}^{i} \boldsymbol{var_{LR}}$	= 6.2382
The sum of the calculated differences $\boldsymbol{Dif_1}$ of the first group of patients	$\boldsymbol{Dif_{1_SUM}} = \sum_{i=1}^{i} \boldsymbol{Dif_1}$	= − 0.4334
The sum of the calculated differences $\boldsymbol{Dif_2}$ of the first group of patients	$\boldsymbol{Dif_{2_SUM}} = \sum_{i=1}^{i} \boldsymbol{Dif_2}$	= 0.4334

The Log Rank statistic is then calculated using the equation (Eq. 31).

$$Log\ Rank = \frac{(Dif_{1_SUM})^2}{var_{LR_SUM}} = \frac{0.1878}{6.2382} = 0.0301 \quad \text{(Eq. 31)}$$

This Log Rank value is compared against the chi-square critical value ($\chi^2{}_{Crit}$) to determine the statistical significance[64]. For this calculation, degrees of freedom (f_{deg}) are calculated using equation (Eq. 32) and known number of rows ($ROWS$) and columns (COL).

$$f_{deg} = (COL - 1) \cdot (ROWS - 1) = (2 - 1) \cdot (7 - 1) = 6 \quad \text{(Eq. 32)}$$

Using a significance level of $\alpha = 0.05$, statistical tables for chi-square or MS Excel®, we find the critical value $\chi^2{}_{Crit} = 12.59$. Since $\chi^2{}_{Crit} > Log\ Rank$, we conclude that the survival distributions of the two groups do not differ significantly, indicating that the inclusion of censored patients does not significantly affect the survival distribution. The Log Rank test confirms H_0, indicating that the effect of censoring is negligible in this context. With larger sample sizes or increased number of events, statistical power may change, potentially revealing significant differences.

Alternative tests, such as the Breslow (Generalized Wilcoxon) test, are more sensitive to early differences, while the Tarone-Ware test balances sensitivity across time frames. In survival data, ties must be handled with adjustments like the Breslow and Peto-Peto modifications. The power of the LR test depends on the number of events rather than the total participants, so studies should be adequately powered with enough events. While MS Excel® can calculate chi-square critical values, dedicated statistical software (e.g., R, SAS, SPSS) offers more robust tools for conducting and interpreting the LR test, especially in complex scenarios.

4 CONCLUSIONS

We presented the application of several statistical methods used in biomedical research, including their manual calculation and interpretation. Descriptive statistics evaluated the time-dependent pain profile in patients after wrist fracture and osteosarcoma resection. We determined the average, standard deviation, and confidence interval of hand muscular strength post-rehabilitation using a dynamometer. Guidelines for advanced statistical methods were re-introduced using examples of real patient groups and hypotheses, with significance levels set for objective ($\alpha = 0.05$), quasi-objective ($\alpha = 0.10$), and subjective parameters ($\alpha = 0.20$). We used the two-sample unpaired T-test to confirm the hypothesis: "Does hand muscular strength change during rehabilitation of patients after wrist fracture?" Correlation analysis described the decrease in serum inflammatory markers post-osteosarcoma resection, slower by 32 ± 2% in females. Appetite increase post-resection was 27 ± 2% faster in males. We identified no significant correlation between age and weight. Non-linear effects on magnetic fields on osteoblast growth were observed during osteosarcoma rehabilitation. Using the paired T-test, we confirmed that two methods for determining serum Ca^{2+} concentrations in post-osteosarcoma patients were equally representative. The F-test showed no significant difference in variances between males and females in recovery outcomes during osteosarcoma remission. To improve the reliance on "Black Box" software and explain for survival analysis, we manually calculated Kaplan-Meier survival curves and the Log Rank test. We confirmed that censoring less than 10 % of patients does not significantly alter survival profiles. The Log Rank test can compare various time-dependent characteristics across patient groups. For future statistical evaluations, advanced fuzzy logic models should be considered. These models, implemented in AI, can incorporate diverse criteria for consistent parameter assessment.

5 ZUSAMMENFASSUNG

Ausgangspunkt: Moderne statistische Modelle sind für eine methodisch fundierte Bewertung von Veränderungen des Gesundheitszustands von Patientinnen und Patienten sowie für eine präzise Prognosestellung

unerlässlich. Im Zeitraum von 2003 bis 2025 wurde in der wissenschaftlichen Gemeinschaft eine Zunahme fehlerhafter statistischer Anwendungen beobachtet. Die Abhängigkeit von Software des Typs „Black Box“, künstlicher Intelligenz oder automatisierten Verfahren ohne gründliches Verständnis der zentralen Parameter führt häufig zu Ungenauigkeiten. Fortschrittliche statistische Software vereinfacht Berechnungen, ersetzt jedoch nicht das Verständnis des verwendeten Modells.
Gruppe: Wir bewerteten Veränderungen des Gesundheitszustands vor und nach Behandlung und Rehabilitation in zwei unterschiedlichen Patientenkollektiven: Patientinnen und Patienten mit Handgelenksfraktur sowie Patientinnen und Patienten mit intramedullärem Osteosarkom. Die muskuläre Erholung nach einer Fraktur erfordert eine präzise Diagnostik, insbesondere den standardisierten Janda Test, der eine zuverlässige Verlaufsbeobachtung ermöglicht. Demgegenüber liegt der Schwerpunkt der Rehabilitation bei Osteosarkom auf dem Vertikalisierungstraining, um die Kontrolle des Körperschwerpunkts zu optimieren und die posturale Muskulatur zu aktivieren. Dieser Ansatz erhält eine maximale Heterogenität der Stichprobe und ermöglicht die Beobachtung von Veränderungen sowohl bei nicht lebensbedrohlichen als auch bei lebensbedrohlichen Erkrankungen.
Methoden: Wir schlagen einen methodologischen Ansatz vor, der eine auf Fuzzy Logik basierende Wertematrix auf verschiedenen Signifikanzniveaus einbezieht, um subjektive, quasi subjektive und objektive Gesundheitsparameter zu bewerten. Daher stellen wir detaillierte manuelle Berechnungen Schritt für Schritt dar, die deskriptive Statistik, Konfidenzintervalle, Korrelationsanalyse, t Tests, F Test, multivariate Analyse, Kaplan Meier Kurven und Log Rank Test umfassen, um methodologische Klarheit sicherzustellen.
Ergebnisse: Unsere Studie zeigte, dass 65 von 600 wissenschaftlichen Arbeiten ungenaue oder nicht optimierte statistische Modelle verwendeten.
Schlussfolgerungen: Ziel dieser Arbeit ist es, einen methodologischen Leitfaden für die korrekte Anwendung fortgeschrittener statistischer Modelle in der wissenschaftlichen Forschung bereitzustellen, insbesondere für Ärztinnen und Ärzte am Beginn ihrer beruflichen Laufbahn. Darüber hinaus soll sie erfahrene Fachleute über mögliche Unklarheiten und Verfahrensschritte bei der Nutzung von Software des Typs „Black Box“ informieren.

6 ACKOWLEDGEMENT

The authors wish to express their sincere gratitude in memoriam to Prof. Anton Guth for his valuable contribution to the preparatory work of this study. We respectfully acknowledge his long-standing support for education and university training in the field of rehabilitation medicine, as well as his important role in building and strengthening the rehabilitation background in Slovakia, Czech Republic, Germany, and Austria. His professional dedication, academic leadership, and lifelong commitment to the development of rehabilitation education and practice remain deeply appreciated and are honoured in this work.

REFERENCES

1. Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S. & Wicherts, J. M. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* **48**, (2016).
2. Wilson, M. On choosing a model for measuring. *MPR-online* **8**, (2003).
3. Masuadi, E. *et al.* Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis. *Cureus* https://doi.org/10.7759/cureus.12639 (2021) doi:10.7759/cureus.12639.
4. Vickers, A. J. *et al.* Guidelines for Reporting of Figures and Tables for Clinical Research in Urology. *Urology* **142**, (2020).
5. Mishra, P., Pandey, C. M., Singh, U. & Gupta, A. Scales of measurement and presentation of statistical data. *Ann. Card. Anaesth.* **21**, (2018).
6. Marshall, G. & Jonker, L. An introduction to descriptive statistics: A review and practical guide. *Radiography* vol. 16 Preprint at https://doi.org/10.1016/j.radi.2010.01.001 (2010).
7. Schober, P. & Vetter, T. R. Analysis of Variance in Medical Research. *Anesth. Analg.* **131**, (2020).
8. Taylor, R. Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography* **6**, (1990).
9. Rich, J. T. *et al.* A practical guide to understanding Kaplan-Meier curves. *Otolaryngology - Head and Neck Surgery* **143**, (2010).
10. Keselman, H. J. *et al.* Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Rev. Educ. Res.* **68**, (1998).
11. Crosby, A. W. *The Measure of Reality*. *The Measure of Reality* (1996). doi:10.1017/cbo9781107050518.
12. Maxim, L. D., Niebo, R. & Utell, M. J. Screening tests: A review with examples. *Inhalation Toxicology* vol. 26 Preprint at https://doi.org/10.3109/08958378.2014.955932 (2014).
13. Sadler, G. R. *et al.* Sample size considerations when groups are the appropriate unit of analysis. *Applied Nursing Research* **20**, (2007).
14. Krousel-Wood, M. A., Chambers, R. B. & Muntner, P. Clinicians' guide to statistics for medical practice and research: Part I. *Ochsner Journal* vol. 6 Preprint at (2006).
15. Wu, X., Wu, B., Sun, J., Qiu, S. & Li, X. A hybrid fuzzy K-harmonic means clustering algorithm. *Appl. Math. Model.* **39**, (2015).

16. Kaliyadan, F. & Kulkarni, V. Types of variables, descriptive statistics, and sample size. *Indian Dermatol. Online J.* **10**, (2019).
17. Strasak, A. M., Zaman, Q., Pfeiffer, K. P., Göbel, G. & Ulmer, H. Statistical errors in medical research - A review of common pitfalls. *Swiss Medical Weekly* vol. 137 Preprint at https://doi.org/10.4414/smw.2007.11587 (2007).
18. Ercan, I. *et al.* Misusage of statistics in medical research. *European Journal of General Medicine* vol. 4 Preprint at https://doi.org/10.29333/ejgm/82507 (2007).
19. Riffenburgh, R. H. & Gillen, D. L. *Statistics in Medicine*. *Statistics in Medicine* (2020). doi:10.1016/B978-0-12-815328-4.00029-2.
20. McDermid, R. Statistics in medicine. *Anaesthesia and Intensive Care Medicine* vol. 22 Preprint at https://doi.org/10.1016/j.mpaic.2021.05.009 (2021).
21. Darling, H. S. Are you confident about your confidence in confidence intervals? *Cancer Research, Statistics, and Treatment* **5**, (2022).
22. Greenland, S. *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* **31**, (2016).
23. Gogtay, N. J. & Thatte, U. M. Principles of correlation analysis. *Journal of Association of Physicians of India* **65**, (2017).
24. Légeret, C., Stienen, Y., Furlano, R. & Köhler, H. Effectivity of treatment for children with functional dyspepsia. *Sci. Rep.* **12**, (2022).
25. Lalkhen, A. G. & McCluskey, A. Statistics V: Introduction to clinical trials and systematic reviews. *Continuing Education in Anaesthesia, Critical Care and Pain* **8**, (2008).
26. Riffenburgh, R. H. Managing Results of Analysis. in *Statistics in Medicine* (2012). doi:10.1016/b978-0-12-384864-2.00015-9.
27. Pallauta, J. D., Arteaga, P., Gea, M. M. & Begué, N. Understanding statistical tables: a survey of research. *Boletin de Estadistica e Investigacion Operativa* **38**, (2022).
28. Müller, H. L. *et al.* Melatonin treatment in obese patients with childhood craniopharyngioma and increased daytime sleepiness. *Cancer Causes and Control* **17**, (2006).
29. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68**, (2015).
30. Xu, M. *et al.* The differences and similarities between two-sample t-test and paired t-test. *Shanghai Arch. Psychiatry* **29**, (2017).
31. Sitter, B. *et al.* Quantification of metabolites in breast cancer patients with different clinical prognosis using HR MAS MR spectroscopy. *NMR Biomed.* **23**, (2010).
32. Kruse, J. A. Clinical Methods: The History, Physical, and Laboratory Examinations. *JAMA: The Journal of the American Medical Association* **264**, (1990).
33. Robertson, W. G., Marshall, R. W. & Walser, M. Calcium measurements in serum and plasma-total and ionized. *Crit. Rev. Clin. Lab. Sci.* **11**, (1979).
34. Zhang, Y., Zhao, F. & Zhang, Y. Exploring the Effectiveness of Nursing Interventions for Postoperative Pain Management in Aortic Dissection: A Fuzzy Logic and Network Analysis. *J. Multidiscip. Healthc.* **18**, (2025).
35. Liu, H. W. Virtual Reality–Assisted Rehabilitation for Patients Undergoing Total Knee Arthroplasty. *Am. J. Phys. Med. Rehabil.* **104**, (2025).
36. Giannou, I., Katsina, M., Dimitriadis, Z., Paras, G. & Besios, T. The effect of hippotherapy on people with multiple sclerosis, a systematic review. *Mult. Scler. Relat. Disord.* **97**, (2025).
37. Hardison, M. E. & Roll, S. C. Mindfulness interventions in physical rehabilitation: A scoping review. *American Journal of Occupational Therapy* **70**, (2016).
38. Galińska, E. Music therapy in neurological rehabilitation settings. *Psychiatr. Pol.* **49**, (2015).
39. Gao, A. *et al.* Rapid identification of Salmonella serovars Enteritidis and Typhimurium using whole cell matrix assisted laser desorption ionization – Time of flight mass spectrometry (MALDI-TOF MS) coupled with multivariate analysis and artificial intelligence. *J. Microbiol. Methods* **213**, (2023).
40. Besse, L. *et al.* Protocol for automated multivariate quantitative-image-based cytometry analysis by fluorescence microscopy of asynchronous adherent cells. *STAR Protoc.* **4**, (2023).
41. Barberis, E. *et al.* Precision Medicine Approaches with Metabolomics and Artificial Intelligence. *International Journal of Molecular Sciences* vol. 23 Preprint at https://doi.org/10.3390/ijms231911269 (2022).
42. Zhang, Z. & Castelló, A. Principal components analysis in clinical studies. *Ann. Transl. Med.* **5**, (2017).
43. Mainous, A. G. Factor analysis as a tool in primary care research. *Fam. Pract.* **10**, (1993).
44. Juanita, S. & Cahyono, R. D. K-MEANS CLUSTERING WITH COMPARISON OF ELBOW AND SILHOUETTE METHODS FOR MEDICINES CLUSTERING BASED ON USER REVIEWS. *Jurnal Teknik Informatika (JUTIF)* **5**, (2024).
45. Sgarro, G. A. *et al.* The Role of BIA Analysis in Osteoporosis Risk Development: Hierarchical Clustering Approach. *Diagnostics* **13**, (2023).
46. Liao, M., Li, Y., Kianifard, F., Obi, E. & Arcona, S. Cluster analysis and its application to healthcare claims data: A study of end-stage renal disease patients who initiated hemodialysis Epidemiology and Health Outcomes. *BMC Nephrol.* **17**, (2016).
47. Pindar Dibal, N. & Akas Abraham, C. On the Application of Linear Discriminant Function to Evaluate Data on Diabetic Patients at the University of Port Harcourt Teaching Hospital, Rivers, Nigeria. *American Journal of Theoretical and Applied Statistics* **9**, (2020).
48. Zhu, F., Gao, J., Yang, J. & Ye, N. Neighborhood linear discriminant analysis. *Pattern Recognit.* **123**, (2022).
49. Koçhan, N., Tutuncu, G. Y., Smyth, G. K., Gandolfo, L. C. & Giner, G. qtQDA: quantile transformed quadratic discriminant analysis for high-dimensional RNA-seq data. *PeerJ* **7**, (2019).
50. Erkartal, H. Ş., Tatlı, M., Secgin, Y., Toy, S. & Duman, B. S. Gender Estimation with Parameters Obtained From the Upper Dental Arcade by Using Machine Learning Algorithms and Artificial Neural Networks. *European Journal of Therapeutics* **29**, (2023).
51. Wang, L. S., Aziz, Z. & Chik, Z. Disposal practice and factors associated with unused medicines in Malaysia: a cross-sectional study. *BMC Public Health* **21**, (2021).
52. Brasier, A. R. *et al.* Predicting intermediate phenotypes in asthma using bronchoalveolar lavage-derived cytokines. *Clin. Transl. Sci.* **3**, (2010).
53. Revilla, L. *et al.* Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis. *PLoS One* **16**, (2021).
54. Gossmann, A., Zille, P., Calhoun, V. & Wang, Y. P. FDR-Corrected Sparse Canonical Correlation Analysis with Applications to Imaging Genomics. *IEEE Trans. Med. Imaging* **37**, (2018).
55. Lee, G. *et al.* Comprehensive Computed Tomography Radiomics Analysis of Lung Adenocarcinoma for Prognostication. *Oncologist* **23**, (2018).

56. Schwartz, L. R., Fernandez, R., Kouyoumjian, S. R., Jones, K. A. & Compton, S. A Randomized Comparison Trial of Case-based Learning versus Human Patient Simulation in Medical Student Education. *Academic Emergency Medicine* **14**, (2007).
57. Goel, M., Khanna, P. & Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.* **1**, (2010).
58. HOLUBOVA, H. Kaplan-Meyer Survival Curves: Simulation Technique. *Scientific Bulletin of the National Academy of Statistics, Accounting and Audit* https://doi.org/10.31767/nasoa.3-4-2021.02 (2021) doi:10.31767/nasoa.3-4-2021.02.
59. Bollschweiler, E. Benefits and limitations of Kaplan-Meier calculations of survival chance in cancer surgery. *Langenbeck's Archives of Surgery* vol. 388 Preprint at https://doi.org/10.1007/s00423-003-0410-6 (2003).
60. Koletsi, D. & Pandis, N. Survival analysis, part 2: Kaplan-Meier method and the log-rank test. *American Journal of Orthodontics and Dentofacial Orthopedics* vol. 152 Preprint at https://doi.org/10.1016/j.ajodo.2017.07.008 (2017).
61. Jager, K. J., Van Dijk, P. C., Zoccali, C. & Dekker, F. W. The analysis of survival data: The Kaplan-Meier method. *Kidney Int.* **74**, (2008).
62. Lin, R. S. & León, L. F. Estimation of treatment effects in weighted log-rank tests. *Contemp. Clin. Trials Commun.* **8**, (2017).
63. D'ascanio, M. *et al.* Age is not the only risk factor in COVID-19: the role of comorbidities and of long staying in residential care homes. *BMC Geriatr.* **21**, (2021).
64. Shen, J. *et al.* Identification of a novel gene signature for the prediction of recurrence in HCC patients by machine learning of genome-wide databases. *Sci. Rep.* **10**, (2020).