Marmara Medical Journal 2017; 30: 101-112
DOI: 10.5472/marumj.344822

**ORIGINAL ARTICLE /** ÖZGÜN ARAŞTIRMA

# Determining the sample size in agreement studies

## Uyum çalışmalarında örneklem büyüklüğünün belirlenmesi

Gülhan TEMEL, Semra ERDOGAN

**ABSTRACT**

**Objective:** Before beginning method comparison studies in clinical researches, all the researchers share a common problem. That is: how to determine the sample size. The aim of this study is to identify the sample size calculation steps for the consistency statistics used in the identification of the agreement between the raters/methods; and to present practical tables belonging to the minimum sample numbers required, before researchers start clinical trials.

**Materials and Methods:** In this study, the steps of sample size calculation have been given for cases where there is no information on neither the population nor the consistency among the raters. Tables have been formed for Cohen Kappa and Intra-class correlation coefficient. Besides, other steps of sample number calculation have been given by utilizing a common formulation used for all consistency statistics by Gwet and practical tables have been presented.

**Results:** When the tables are studied, no matter what the importance level and the test power is, the sample number increases while inconsistency rate between the two raters increases up to 0.50; and the sample number shows a symmetrical decrease while inconsistency rate between the two raters shows an increase from 0.50 through 1. Moreover, as the consistency value between the raters rise, no matter what the test power and the importance level is, the sample size to be included in the study decreases in direct proportion.

**Conclusion:** Before beginning a research study, with the exact determination of the minimum number of samples enough for the design of the study and the state of the final variable, besides proving reliability of the results of the study, sampling waste will also be prevented.

**Keywords:** Method comparison, Agreement between the raters, Sample size

**ÖZ**

**Amaç:** Klinik araştırmalarda metot karşılaştırması çalışmalarına başlamadan önce tüm araştırmacıların problem yaşadığı şey ne kadar örneklem büyüklüğü ile çalışılmasıdır. Bu çalışmanın amacı, değerlendiriciler / yöntemler arasındaki uyumun belirlenmesinde kullanılan uyum istatistikleri için örneklem büyüklüğünün hesaplanma adımlarını tanımlamak, klinik çalışmalar için araştırmacılara araştırmaya başlamadan önce gerekli olan minimum örneklem sayılarına ait pratik tablolar sunmaktır.

**Gereçler ve Yöntemler:** Bu çalışmada, populasyona ait bir bilgi olmadığı durumda ve değerlendiriciler arası uyum bilindiğinde örneklem büyüklüğünün hesaplama adımları verilmiştir. Cohen Kappa ve Sınıf içi korelasyon katsayısı için tablolar oluşturulmuştur. Ayrıca Gwet tarafından tüm uyum istatistikleri için kullanılabilecek ortak bir formülasyondan yararlanılarak da örneklem büyüklüğü hesaplama adımları verilmiş ve pratik tablolar sunulmuştur.

**Bulgular:** Tablolar incelendiğinde, önem seviyesi ve testin gücü ne olursa olsun iki değerlendirici arasındaki uyumsuzluğun oranı 0.50'ye kadar artış gösterirken örneklem büyüklüğü de artmakta, 0.50'den 1'e doğru artış gösterirken simetrik olarak bir azalış göstermektedir. Bunun yanı sıra, değerlendiriciler arasındaki uyum değeri arttıkça testin gücü ve önem seviyesi ne olursa olsun doğru orantılı olarak çalışmaya dahil edilecek olan örneklem sayısı da azalmaktadır.

**Sonuç:** Bir araştırma çalışmasının başlangıcında, çalışmanın tasarımına ve sonuç değişkeninin durumuna uygun olan yeterli minimum örneklem sayısının doğru olarak belirlenmesi ile, çalışma sonuçlarının güvenilirliği sağlanmış olmasının yanında, örneklem israfının da önüne geçilmiş olacaktır.

**Anahtar kelimeler:** Metot karşılaştırması, Değerlendiriciler arası uyum, Örneklem büyüklüğü

## Introduction

Method comparison in clinical researches is, to assess whether the two different techniques handled, are in agreement or not, and whether the technique that can be used as an alternative to reference technique is valid

Gülhan Temel, Semra Erdoğan (✉)
*Department of Biostatistics and Medical Informatics, School of Medicine, Mersin University, Mersin, Turkey*
*e-mail: semraerdogann@gmail.com*

**Submitted / Gönderilme:** 05.03.2017 **Accepted/Kabul:** 13.04.2017

or even superior or not. As the data obtained are based on a measurement, the measurement done or the new technique developed have to be shown to be valid and reliable to be used. Reliability is not only used in the comparison of measurement methods but also in the test of the compatibility between the measurements obtained from the repeated measurements of a single measurement method (the rater) or from two or more raters (method). The statistical method to be applied for the identification of the degree of reproducibility of the measurements, the agreement between the measurements between the methods or measurements taken by more than one raters depends on what kind of a variable the result of measurement has been expressed and the number of raters [1].

Agreement studies is quite common in studies in medical field as well as educational sciences. Besides, the most frequent problem all the researchers face with before starting such studies is required sample size they should study with; because the detection of the sampling size is one of the most important and even most difficult steps in the planning of clinical studies. Studying with enough samples is quite important in scientific, economical and ethical terms. Studying on a sample size large enough is the most important factor in guaranteeing the validity and reliability of the findings obtained from the study scientifically. In scientific studies, while studying with less than enough samples in number would decrease the power of the results of the study, studying with more than enough samples would lead to a futile effort and resource waste. Besides, as studying with more than necessary number of samples would lead to the exposure of many individuals to unnecessary harmful factors, that would create ethically inappropriate results. Many researchers and scientific publications get help from many guides and standards appropriate to the studies and developed accordingly in order to increase the quality and the reliability of the clinical studies. These kind of guides' having the question of "How the sampling size is determined" in their method parts show how greatly important the determination of the sampling size and power in clinical studies is [2, 3].

Before the determination of sample size, it is necessary that the population is known well and the effect size representing how much type 1 error, the test power and the estimated value obtained as a result of the study would digress is determined.

In agreement studies, if there are preexisting studies, sample size can be calculated with the help of the agreement statistics between the raters in these studies. Sometimes, while calculating the sample size, there may not be any information about the population or may not be a preexisting study. In this case, the researcher is advised to do a pilot study. However, if the researcher does not have enough time to do a pilot study, only when the type 1 error, the power of the test and a significance difference between the two raters (effect size) is known can the sample size be calculated. The formulations of power and sample size show difference for each clinical study. The aim of this study is to describe the sample size calculation steps for agreement statistics used in the determination of the agreement between the raters/methods and to present practical tables about the minimum sample size required for researchers before they start the research in clinical studies.

## Materials and Methods

### *Agreement Statistics*

Assume that the result variable for a disease interested by two different raters like A and B is determined as patient (+) and healthy (-). This case is shown with a contingency table as in Table I.

In Table I, while the diagonal values show that the measurements belonging to both evaluators are in agreement, the measurement results except the diagonal values show disagreement. In other words, while $d_{11}$ shows the number of classifications in the category "+" by both raters and $d_{00}$ shows the number of classifications in the category of "–" by both raters, $d_{01}$ and $d_{10}$ shows the number of classifications in the situations where both raters disagree [4]. Agreement probability is calculated by utilizing the sum of the number

Table I: Cross table belonging to a state of two raters and two categories.

| | | **Rater A** | | |
|---|---|---|---|---|
| | | - | + | Total |
| **Rater B** | - | $d_{00}$ | $d_{01}$ | |
| | + | $d_{10}$ | $d_{11}$ | |
| | Total | | | N |

of same classification (+/- categories) for both raters, also disagreement probability is calculated by utilizing the sum of the number of different classification (+/- categories) for raters. The probabilities of both raters being in agreement ($\pi_A$) and disagreement ($\pi_D$) is given in Equation 1.

$$\pi_A = \frac{d_{00} + d_{11}}{N} \text{ and } \pi_D = \frac{d_{01} + d_{10}}{N} \quad (1)$$

If the measurements belonging to both raters and the agreement coefficient between these raters is not known but the two raters' being in disagreement is possible, the sample size required for these cases can be calculated by using Equation 2. Here $\pi_D$ shows the disagreement probability of two raters and $W_D$ shows type II error (β). The power of the test is expressed as (1-β). In the case that the power of the test is 80 %, type II error value will be 0.20. The $Z^2_{1-\alpha/2}$ in Equation 2 gives the probability values in standard normal distribution table belonging to the significance levels (Type I error, alpha). The probability value belonging to the normal distribution table for 0.001 significance level is 3.2905 and 2.5758 for 0.01 significance level and 1.96 for 0.05 significance level [5].

$$n = \frac{4\pi_D(1-\pi_D)Z^2_{1-\alpha/2}}{W_D^2} \quad (2)$$

***Cohen Kappa Statistics***

Kappa statistics have been developed in 1960 by Cohen in order to evaluate the agreement between the two raters and have been formulized as in Equation 3 [6].

$$\kappa = \frac{\pi_A - \pi_E}{1 - \pi_E} \quad (3)$$

$\pi_A$ shows the rate of both raters' being in perfect agreement and is calculated as $(d_{00}+d_{11})/N$. $\pi_E$ shows the expected rate of chance agreement probability and is obtained from the sum of the negative and positive agreements. Negative agreement is calculated by utilizing the sum of row and column of negative results for two raters, positive agreement is calculated by utilizing the sum of row and column of positive results. Negative agreement is formulized by using Equation 4 and positive agreement is formulized by using Equality 5 and the proportion of chance agreement is formulized as in Equality 6 [5].

$$Negatif\ Agreement = \frac{(d_{00} + d_{01})(d_{00} + d_{10})}{N^2} \quad (4)$$

$$Pozitif\ Agreement = \frac{(d_{10} + d_{11})(d_{01} + d_{11})}{N^2} \quad (5)$$

$$\pi_E = \frac{(d_{00} + d_{01})(d_{00} + d_{10})}{N^2} + \frac{(d_{10} + d_{11})(d_{01} + d_{11})}{N^2} \quad (6)$$

Kappa coefficient takes a value between -1 and +1 but practically, a value between 0 and 1 is interested. So, the interpretable interval of Kappa coefficient is 0 and +1, and its being smaller than 0 (negative) does not have a meaning in terms of reliability. While Kappa coefficient's taking the value of 1 is said to be a perfect agreement, its taking the value 0 gives the result that there is no agreement and the decisions of the two raters are completely different [1]. In some sources, there are different classifications about the strength of the Kappa coefficient (agreement degree). These classifications may change on the topic that is studied. Generally, in the studies, as the limit value for Kappa coefficient, 0.20 is said to be poor, 0.21-0.40 is below the moderate, 0.41-0.60 is moderate, 0.61-0.80 is good and 0.81-1.00 is said to be in perfect agreement [7].

The quite common agreement statistic in the literature is Cohen Kappa statistics. However, it has been put forward that this agreement statistic is affected from sensitivity, specificity and prevalence and that it has to be more carefully handled while being used in reliability and agreement studies [8].

The agreement between the raters can be calculated with the help of Cohen Kappa statistic. In this case, before beginning the study, when the agreement coefficient between the raters is known, minimum necessary sample size ($m_k$) can be calculated as in Equation 7. In the equation, $\pi_{Dis}$ shows the probability of disagreement and $W_D$ shows the type II error. $Z^2_{1-\alpha/2}$ gives the probability values in standard distribution table belonging to significance levels (Type I error) [5].

$$m_\kappa = 4\frac{(1-\kappa)}{W_D^2}\left((1-\kappa)(1-2\kappa) + \frac{\kappa(2-\kappa)}{2\pi_D(1-\pi_D)}\right)Z^2_{1-\alpha/2} \quad (7)$$

***Intra-class Correlation Coefficient* (*ICC*)**

If our result variable is not categorical but in a continuous structure, Intra-class correlation coefficient is used for the agreement between the raters and when there are 2 or more raters, ICC value is formulized as in Equation 8. In the equation, $\sigma_B$ represents the standard deviation between the raters and $\sigma_W$ indicates the standard deviation within the raters [9].

(4)

$$\rho = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2} \quad (8)$$

The acceptable level of ICC shows difference according to the characteristics of data and the subject studied and the aim. When the agreement between the raters is considering, ICC is expected to be minimum 0.70. It can be said that the agreement between the raters is "perfect" if the ICC is taking a value between 0.95 and 1.00, "high" if it takes a value between 0.85 and 0.94, "moderate" level if it takes a value between 0.70 and 0.84 and no agreement at all if it takes a value below 0.70 [10, 11].

According to the confidence interval approach, in a case where there are k number of raters independent from each other, before starting the study, the minimum necessary sample size ($m_{ICC}$) is calculated as in Equation 9 when the agreement correlation between the raters is known. In the equation, $W_D$ shows type II error, $Z^2_{1-\alpha/2}$ shows the probability values in the standard normal distribution table belonging to the significance levels (Type I error), $\rho_{Plan}$ shows the ICC [5]. And it is obtained from a prior research or an expert opinion. The sample size that is determined for ICC is valid for three ANOVA models of ICC and for measurement reliability (Consistency, absolute agreement) [12].

$$m_{ICC} = 1 + \frac{8\, Z^2_{1-\alpha/2} \left(1-\rho_{Plan}\right)^2 \left[1+(k-1)\rho_{Plan}\right]^2}{k(k-1)W_D^2} \quad (9)$$

### *Alternative Sample Size Formulation for Gwet's Agreement Statistics*

Gwet's agreement coefficient (AC1), has been put forward in 2001 by Gwet and is calculated as (Equation 10) [6,13].

$$AC1 = \gamma = \frac{\pi_A - \pi_E(\gamma)}{1 - \pi_E(\gamma)} \quad (10)$$

In the equation, the proportion of chance agreement is calculated as $\pi_E(\gamma) = 2P_1(1-P_1)$. The $P_1$ probability in the formula is calculated as in Equation 11 [6,13].

$$P_1 = \frac{(d_{00} + d_{10} + d_{00} + d_{01})/2}{N} \quad (11)$$

The AC1 statistic put forward by Gwet is said to be not affected from sensitivity, specificity and prevalence values compared to Cohen's Kappa statistic and to show a better performance [8]. Besides, if the prevalence value is known and matters for the study, the use of Gwet's AC1 statistics is advised to the researchers [14].

While the reliability coefficient between the real raters is obtained based on the entire population, the estimated reliability coefficient between the raters is obtained from the sample. According to Gwet, for the reliability coefficients between the raters to be valid, it has to be more than 20 % of the true value (the value obtained based on the population). Here, the value 20 % is taken arbitrary and can be changed by the researchers. Gwet argues that the sample size is affected from this arbitrary value in the reliability studies. Based on that, he suggests the formulation in Equation 12 for the calculation of necessary minimum sample size in agreement studies. In this formulation, N shows the sample size belonging to the population, r shows the relative error (the difference between the true value obtained from the population and the estimated value obtained from the sample, effect size), $\pi_A$ shows the overall agreement probability and $\pi_E$ shows the chance-agreement probability [15, 16].

$$n = \frac{n^*}{1 + \frac{n^*}{N}}, \quad n^* = \frac{1}{r^2\left(\pi_A - \pi_E\right)^2} \quad (12)$$

## Results

Determination of sample size is one of the most important and even a difficult step in the planning and designing of clinical studies. This is why the suggested sample size formulations in the determination of minimum sample size enough to the researcher at the beginning of a study in agreement studies are calculated under different conditions and presented in tables. With this aim, a macro has been written in Excel for each formulation and the results obtained have been put into tables. Besides, with the help of the Demo version of SPSS 21 statistic packet program, the graphics for Tables 2-5 have been obtained.

In this study, the calculation steps for sample sizes have been given when there is no information about the population and the agreement between the raters is known. In this study, the tables have been prepared only for Cohen Kappa and ICC. Besides, sample size calculation steps have been given with the help of a common formulation that can

be used for all agreement statistics by Gwet and practical tables have been presented.

To calculate enough sample size according to the situation where the measurements belonging to the two raters and the agreement coefficient between these raters is known but it is possible that these two raters are in disagreement, with the help of Equation 2, the minimum necessary sample size has been calculated in 0.001, 0.01 and 0.05 significance levels, various power levels (95 %, 90 % and 80 %) and for 16 different disagreement rates (0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.60, 0.70, 0.80, 0.90 and 1) and has been given in Table II. The graphic belonging to Table II is as in Figure 1. When Table II and Figure 1 are considered, while the rate of disagreement between the two raters shows an increase up to 0.50, size sample also increases and shows a symmetrical decrease after 0.50. When the disagreement probability is 100 %, the sample size is calculated as 0. As the test power increases, for the agreement between the raters to be significance, it is necessary that more samples are studied.

With the help of the Equation belonging to Kappa statistic suggested by Cohen to calculate the agreement between the raters, the minimum necessary sample size has been calculated for three different type 1 errors (0.001, 0.01, and 0.05), three different test powers (95 %, 90 % and 80 %), 6 different disagreement rates (0.05, 0.10, 0.20, 0.30, 0.40, 0.50) and 9 different Kappa statistic values (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80 and 0.90) and has been given in Table III. Besides, calculations for disagreement rates 0.60, 0.70, 0.80, 0.90 and 0.95 have also been done but 0.05 has given the same results with 0.95, 0.10 with 0.90, 0.20 with 0.80, 0.30 with 0.70 and 0.40 with 0.60. This is why the results belonging to only 6 different disagreement rates have been written in the table. Besides, it has been observed that, in all possible disagreement rates, all type I and type II errors get the value 0 when the Kappa statistic is "1" and that it always gives the same results when Kappa statistic is "0". The graphic belonging to Table III is given in Figure 2. When Table III and Figure 2 are considered, no matter what the disagreement rates between the raters, type I error and test power are, enough

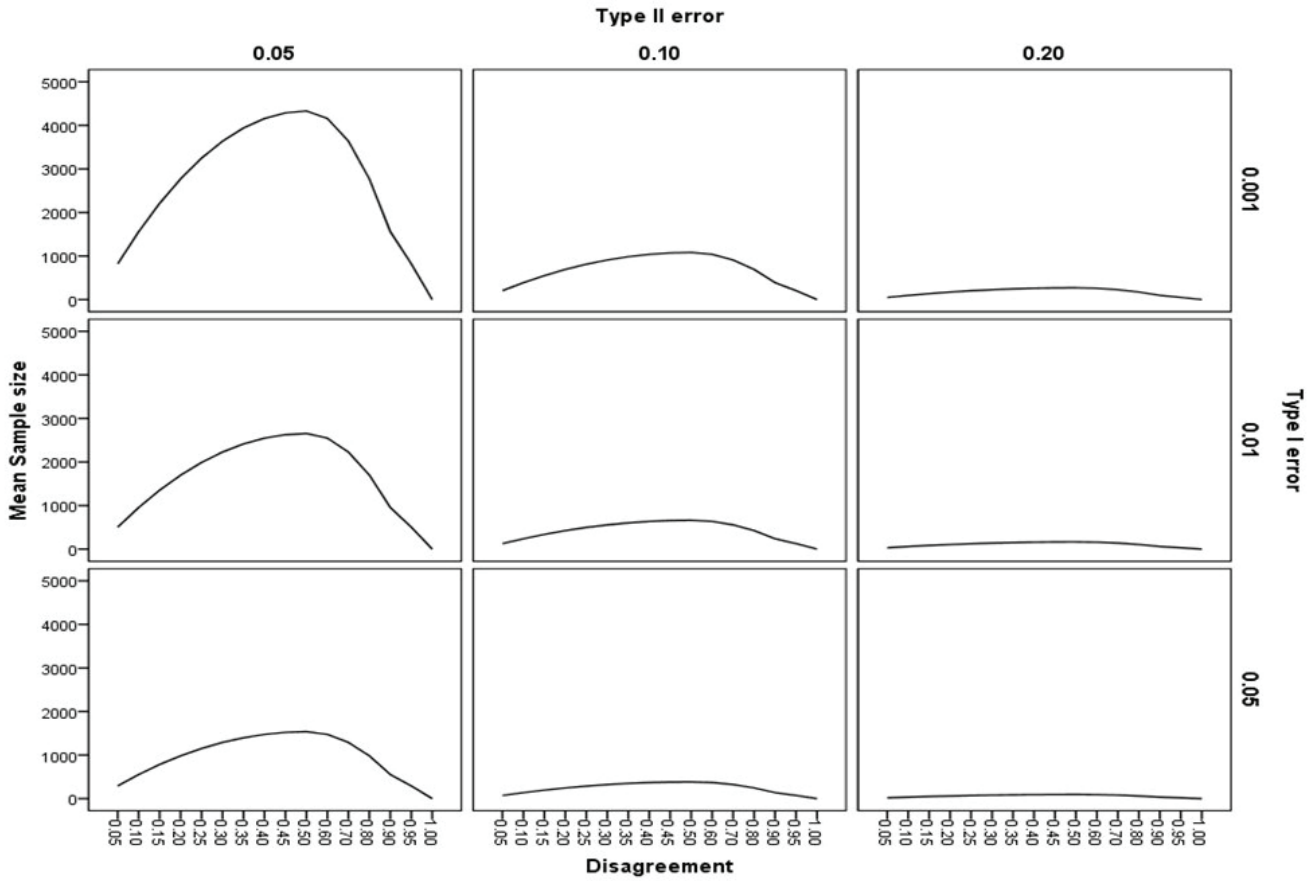


Figure 1: The necessary sample sizes according to the disagreement rates between the two raters, Type I and Type II errors.

Table II: The necessary sample sizes according to the disagreement rates between the two raters, Type I and Type II errors.

| | alfa=0.001 | | | alfa=0.01 | | | alfa=0.05 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\pi_D$ | β= 0.05 | β= 0.10 | β= 0.20 | β= 0.05 | β= 0.10 | β= 0.20 | β= 0.05 | β= 0.10 | β= 0.20 |
| **0.05** | 823 | 206 | 51 | 504 | 126 | 32 | 292 | 73 | 18 |
| **0.10** | 1559 | 390 | 97 | 955 | 239 | 60 | 553 | 138 | 35 |
| **0.15** | 2209 | 552 | 138 | 1353 | 338 | 85 | 784 | 196 | 50 |
| **0.20** | 2772 | 693 | 173 | 1698 | 425 | 106 | 983 | 246 | 61 |
| **0.25** | 3248 | 812 | 203 | 1990 | 498 | 124 | 1152 | 288 | 72 |
| **0.30** | 3638 | 910 | 227 | 2229 | 557 | 139 | 1291 | 323 | 81 |
| **0.35** | 3941 | 985 | 246 | 2415 | 604 | 151 | 1398 | 350 | 87 |
| **0.40** | 4158 | 1039 | 260 | 2548 | 637 | 159 | 1475 | 369 | 92 |
| **0.45** | 4288 | 1072 | 268 | 2627 | 657 | 164 | 1521 | 380 | 95 |
| **0.50** | 4331 | 1083 | 271 | 2654 | 663 | 166 | 1537 | 384 | 96 |
| **0.60** | 4158 | 1039 | 260 | 2548 | 637 | 159 | 1475 | 369 | 92 |
| **0.70** | 3638 | 910 | 227 | 2229 | 557 | 139 | 1291 | 323 | 81 |
| **0.80** | 2772 | 693 | 173 | 1698 | 425 | 106 | 983 | 246 | 61 |
| **0.90** | 1559 | 390 | 97 | 955 | 239 | 60 | 553 | 138 | 35 |
| **0.95** | 823 | 206 | 51 | 504 | 126 | 32 | 292 | 73 | 18 |
| **1.00** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**$\pi_D$:Disagreement rates; alfa: Type I error; β:Type II error**

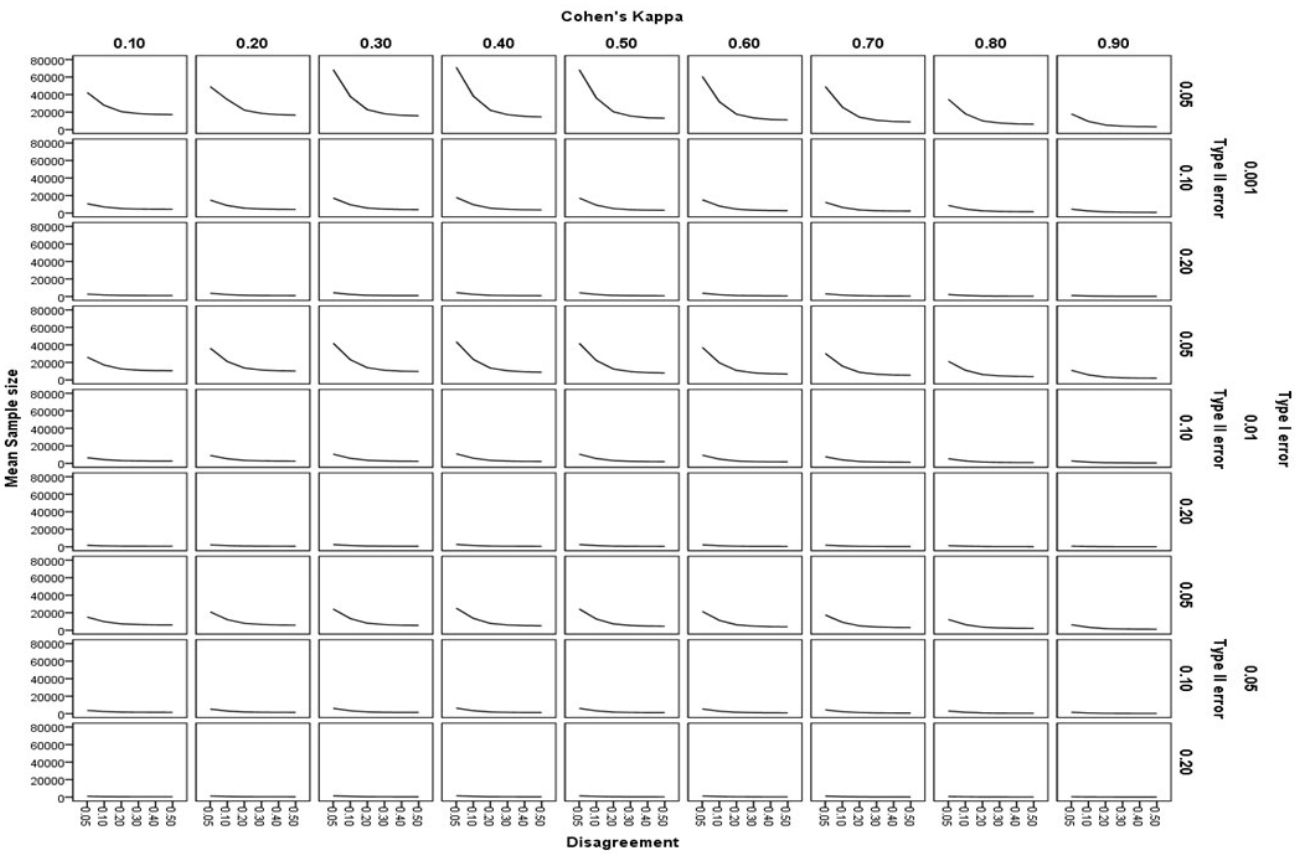


Figure 2: The necessary sample sizes according to Type I and Type II errors given by using Cohen Kappa.

sample size increases while Kappa statistic is rising up to 0.50 and the enough sample size decreases when the Kappa statistic shows an increase from 0.50 to 1. Besides, while the disagreement rate between the raters shows an increase up to 0.50, the enough sample size also decreases no matter what the Kappa statistic value, type I error and test power are; the disagreement rate between the raters shows an increase from 0.50 to 0.95, the enough sample size shows a symmetrical increase no matter what the Kappa statistic value, type I error and test power are.

Table III: The necessary sample sizes according to Type I and Type II errors given by using Cohen Kappa.

| | | alfa=0.001 | | | alfa=0.01 | | | alfa=0.05 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\pi_D$ | Kappa | β= 0.05 | β= 0.10 | β= 0.20 | β= 0.05 | β= 0.10 | β= 0.20 | β= 0.05 | β= 0.10 | β= 0.20 |
| 0.05 | 0.10 | 42409 | 10602 | 2651 | 25987 | 6497 | 1624 | 15047 | 3762 | 940 |
| | 0.20 | 49171 | 14793 | 3698 | 36258 | 9065 | 2266 | 20994 | 5249 | 1312 |
| | 0.30 | 68497 | 17124 | 4281 | 41973 | 10493 | 2623 | 24303 | 6076 | 1519 |
| | 0.40 | 71272 | 17818 | 4455 | 43674 | 10918 | 2730 | 25288 | 6322 | 1580 |
| | 0.50 | 68384 | 17096 | 4274 | 41904 | 10476 | 2619 | 24263 | 6066 | 1516 |
| | 0.60 | 60717 | 15179 | 3795 | 37206 | 9301 | 2325 | 21543 | 5386 | 1346 |
| | 0.70 | 49160 | 12290 | 3072 | 30124 | 7531 | 1882 | 17442 | 4360 | 1090 |
| | 0.80 | 34597 | 8649 | 2162 | 21200 | 5300 | 1325 | 12275 | 3069 | 767 |
| | 0.90 | 17915 | 4479 | 1120 | 10978 | 2744 | 686 | 6356 | 1589 | 397 |
| 0.10 | 0.10 | 27683 | 6921 | 1730 | 16964 | 4241 | 1060 | 9822 | 2456 | 614 |
| | 0.20 | 34370 | 8593 | 2148 | 21061 | 5265 | 1316 | 12195 | 3049 | 762 |
| | 0.30 | 37754 | 9439 | 2360 | 23135 | 5784 | 1446 | 13395 | 3349 | 837 |
| | 0.40 | 38205 | 9551 | 2388 | 23411 | 5853 | 1463 | 13555 | 3389 | 847 |
| | 0.50 | 36091 | 9023 | 2256 | 22116 | 5529 | 1382 | 12805 | 3201 | 800 |
| | 0.60 | 31783 | 7946 | 1986 | 19476 | 4869 | 1217 | 11277 | 2819 | 705 |
| | 0.70 | 25651 | 6413 | 1603 | 15718 | 3930 | 982 | 9101 | 2275 | 569 |
| | 0.80 | 18063 | 4516 | 1129 | 11069 | 2767 | 692 | 6409 | 1602 | 401 |
| | 0.90 | 9390 | 2347 | 587 | 5754 | 1438 | 360 | 3331 | 833 | 208 |
| 0.20 | 0.10 | 20483 | 5121 | 1280 | 12552 | 3138 | 784 | 7268 | 1817 | 454 |
| | 0.20 | 22244 | 5561 | 1390 | 13630 | 3408 | 852 | 7892 | 1973 | 493 |
| | 0.30 | 22722 | 5681 | 1420 | 13924 | 3481 | 870 | 8062 | 2015 | 504 |
| | 0.40 | 22036 | 5509 | 1377 | 13503 | 3376 | 844 | 7818 | 1954 | 489 |
| | 0.50 | 20301 | 5075 | 1269 | 12440 | 3110 | 778 | 7203 | 1801 | 451 |
| | 0.60 | 17636 | 4409 | 1102 | 10806 | 2702 | 675 | 6257 | 1564 | 391 |
| | 0.70 | 14156 | 3539 | 885 | 8674 | 2169 | 542 | 5023 | 1256 | 314 |
| | 0.80 | 9979 | 2495 | 624 | 6115 | 1529 | 382 | 3540 | 885 | 221 |
| | 0.90 | 5221 | 1305 | 326 | 3199 | 800 | 200 | 1852 | 463 | 116 |
| 0.30 | 0.10 | 18279 | 4570 | 1142 | 11201 | 2800 | 700 | 6485 | 1621 | 405 |
| | 0.20 | 18532 | 4633 | 1158 | 11356 | 2839 | 710 | 6575 | 1644 | 411 |
| | 0.30 | 18121 | 4530 | 1133 | 11104 | 2776 | 694 | 6429 | 1607 | 402 |
| | 0.40 | 17086 | 4272 | 1068 | 10470 | 2618 | 654 | 6062 | 1516 | 379 |
| | 0.50 | 15468 | 3867 | 967 | 9478 | 2370 | 592 | 5488 | 1372 | 343 |
| | 0.60 | 13305 | 3326 | 832 | 8153 | 2038 | 509 | 4721 | 1180 | 295 |
| | 0.70 | 10637 | 2659 | 665 | 6518 | 1629 | 407 | 3774 | 943 | 236 |
| | 0.80 | 7504 | 1876 | 469 | 4598 | 1150 | 287 | 2662 | 666 | 166 |
| | 0.90 | 3945 | 986 | 247 | 2417 | 604 | 151 | 1400 | 350 | 87 |
| 0.40 | 0.10 | 17397 | 4349 | 1087 | 10661 | 2665 | 666 | 6173 | 1543 | 386 |
| | 0.20 | 17047 | 4262 | 1065 | 10446 | 2611 | 653 | 6048 | 1512 | 378 |
| | 0.30 | 16280 | 4070 | 1018 | 9976 | 2494 | 623 | 5776 | 1444 | 361 |
| | 0.40 | 15106 | 3777 | 944 | 9257 | 2314 | 578 | 5360 | 1340 | 335 |
| | 0.50 | 13534 | 3384 | 846 | 8293 | 2073 | 518 | 4802 | 1201 | 300 |
| | 0.60 | 11572 | 2893 | 723 | 7091 | 1773 | 443 | 4106 | 1026 | 256 |
| | 0.70 | 9229 | 2307 | 577 | 5655 | 1414 | 353 | 3275 | 819 | 205 |
| | 0.80 | 6514 | 1628 | 407 | 3991 | 998 | 249 | 2311 | 578 | 144 |
| | 0.90 | 3434 | 859 | 215 | 2105 | 526 | 132 | 1219 | 305 | 76 |
| 0.50 | 0.10 | 17151 | 4288 | 1072 | 10509 | 2627 | 657 | 6085 | 1521 | 380 |
| | 0.20 | 16631 | 4158 | 1039 | 10191 | 2548 | 637 | 5901 | 1475 | 369 |
| | 0.30 | 15765 | 3941 | 985 | 9660 | 2415 | 604 | 5593 | 1398 | 350 |
| | 0.40 | 14552 | 3638 | 910 | 8917 | 2229 | 557 | 5163 | 1291 | 323 |
| | 0.50 | 12993 | 3248 | 812 | 7962 | 1990 | 498 | 4609 | 1152 | 288 |
| | 0.60 | 11087 | 2772 | 693 | 6794 | 1698 | 425 | 3934 | 983 | 246 |
| | 0.70 | 8835 | 2209 | 552 | 5414 | 1353 | 338 | 3135 | 784 | 196 |
| | 0.80 | 6237 | 1559 | 390 | 3822 | 955 | 239 | 2213 | 553 | 138 |
| | 0.90 | 3292 | 823 | 206 | 2017 | 504 | 126 | 1168 | 292 | 73 |

**$\pi_D$:Disagreement probability; alfa: Type I error; β:Type II error**

With the help of the Equation belonging to ICC, the necessary minimum sample sizes have been calculated for three different type I error (0.001, 0.01 and 0.05), three different test power (95 %, 90 % and 80 %), 8 different ICC values (0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90 and 0.95) and 3 different number of raters (2, 3 and 4) and have been given in Table IV. The graphic belonging to the Table IV is as in Figure 3. The numbers belonging to the samples sizes given in tables express the sample size necessary for only one rater. When Table IV and Figure 3 are examined, the sample number to be included in the study decreases directly proportionally regardless of the test power and significance level. When the number of raters is 2, more sample is needed to be studied compared to 3 or 4 raters.

With the sample size formulation for all possible agreement statistics by Gwet, the minimum sample sizes to be pulled from the population have been calculated in five different relative errors (effect size) (0.10, 0.20, 0.30, 0.40 and 0.50), the value differences between 10 different agreement probabilities (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 and 1) and in five different sample sizes belonging to the population (30, 100, 500, 1000 and 10000) and have been given in Table V. The graphical display of Table V is as in Figure 4. When Table V and Figure 4 are examined, there is an inverse proportion between the relative error and sample size. As effect size increases, the enough sample size decreases.

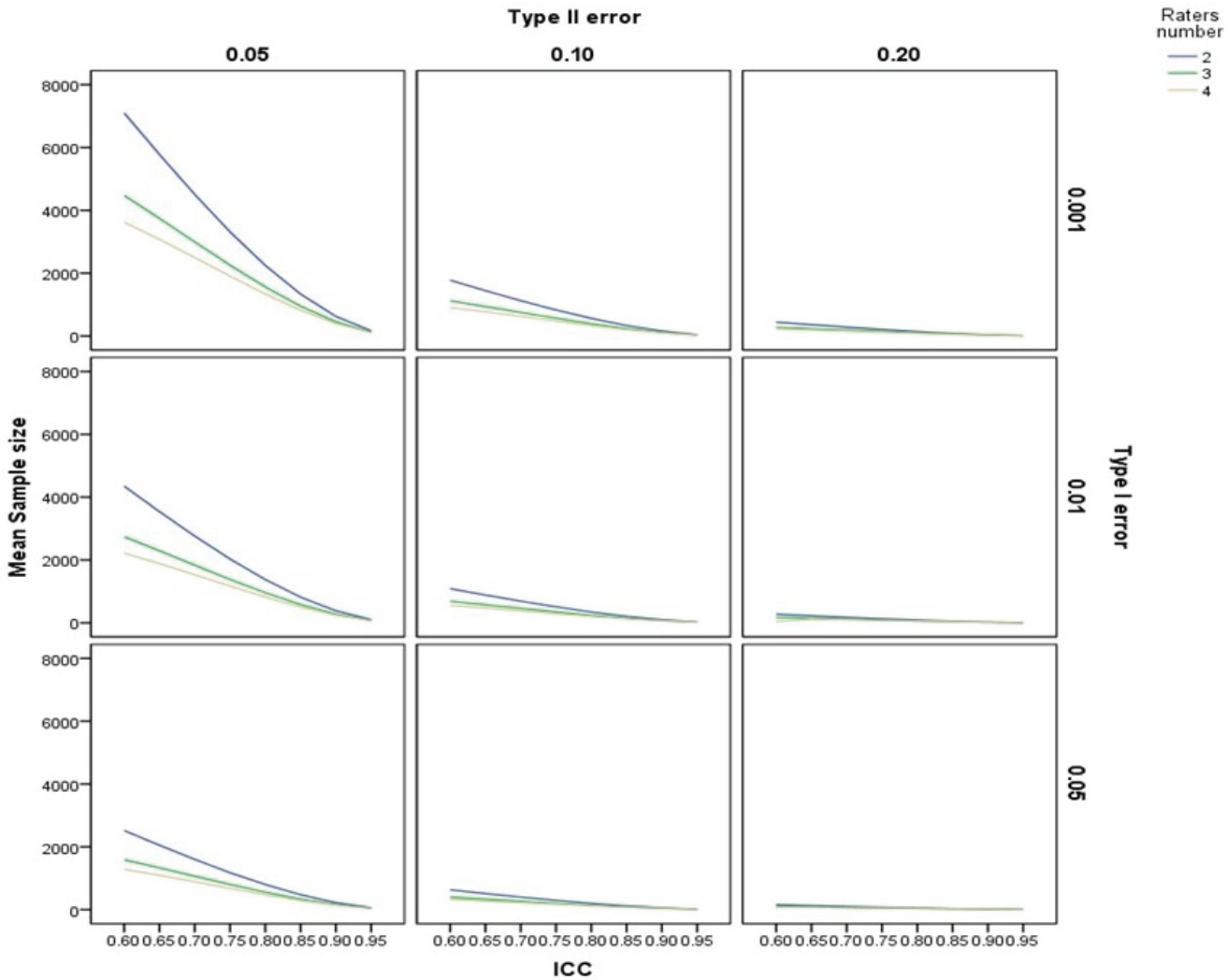


Figure 3: The necessary sample sizes according to Type I and Type II errors given for the agreement between the raters by using intra-class correlation coefficient (ICC).

Table IV: The necessary sample sizes according to Type I and Type II errors given for the agreement between the raters by using intra-class correlation coefficient (ICC).

| | | alfa=0.001 | | | alfa=0.01 | | | alfa=0.05 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **k** | **ICC** | **β=0.05** | **β=0.10** | **β=0.20** | **β=0.05** | **β=0.10** | **β=0.20** | **β=0.05** | **β=0.10** | **β=0.20** |
| **2** | **0.60** | 7097 | 1775 | 444 | 4349 | 1088 | 273 | 2519 | 630 | 158 |
| | **0.65** | 5779 | 1445 | 362 | 3541 | 886 | 222 | 2051 | 513 | 129 |
| | **0.70** | 4507 | 1127 | 283 | 2762 | 691 | 174 | 1600 | 401 | 101 |
| | **0.75** | 3317 | 830 | 208 | 2033 | 509 | 128 | 1177 | 295 | 75 |
| | **0.80** | 2246 | 562 | 141 | 1377 | 345 | 87 | 798 | 200 | 51 |
| | **0.85** | 1335 | 335 | 84 | 818 | 205 | 52 | 474 | 119 | 31 |
| | **0.90** | 626 | 157 | 40 | 384 | 97 | 25 | 223 | 56 | 15 |
| | **0.95** | 166 | 42 | 11 | 102 | 26 | 7 | 59 | 16 | 5 |
| **3** | **0.60** | 4473 | 1119 | 280 | 2741 | 686 | 172 | 1588 | 398 | 100 |
| | **0.65** | 3743 | 937 | 235 | 2294 | 574 | 144 | 1329 | 333 | 84 |
| | **0.70** | 2995 | 749 | 188 | 1835 | 460 | 116 | 1063 | 267 | 67 |
| | **0.75** | 2257 | 565 | 142 | 1383 | 347 | 87 | 801 | 201 | 51 |
| | **0.80** | 1562 | 391 | 99 | 958 | 240 | 61 | 555 | 140 | 36 |
| | **0.85** | 948 | 238 | 60 | 581 | 146 | 37 | 337 | 85 | 22 |
| | **0.90** | 454 | 114 | 29 | 278 | 70 | 18 | 162 | 41 | 11 |
| | **0.95** | 122 | 31 | 9 | 75 | 20 | 6 | 44 | 12 | 4 |
| **4** | **0.60** | 3623 | 906 | 227 | 2220 | 556 | 40 | 1286 | 322 | 81 |
| | **0.65** | 3079 | 771 | 193 | 1887 | 473 | 119 | 1093 | 274 | 69 |
| | **0.70** | 2498 | 625 | 157 | 1531 | 384 | 97 | 887 | 223 | 56 |
| | **0.75** | 1907 | 478 | 120 | 1169 | 293 | 74 | 677 | 170 | 43 |
| | **0.80** | 1336 | 335 | 84 | 819 | 206 | 52 | 475 | 119 | 31 |
| | **0.85** | 820 | 206 | 52 | 503 | 126 | 32 | 291 | 74 | 19 |
| | **0.90** | 396 | 100 | 26 | 243 | 62 | 16 | 141 | 36 | 10 |
| | **0.95** | 108 | 28 | 8 | 67 | 17 | 5 | 39 | 10 | 3 |

**k: Number of raters; ICC: Intra-class correlation coefficient; alfa: Type I error; β:Type II error**

Effect size is a very important concept in the determination of sample size. Effect size shows difference according to the studies. Effect size is sometimes determined with literature scan with the help of the studies done before, and sometimes it can be determined with a pilot study in the absence of a study done before. Sometimes, none of them happen, in such a case, according to the advice of Cohen, the effect size is determined with the help of predetermined 0.20 low effect, 0.50 middle level impact and 0.80 high impact or other pieces of advice [6, 12]. Based on these, the sample size for different agreement probabilities from different size populations have been calculated and as is Table V.

### *Strengths and Limitations of the Study*

#### *Strengths*

The most important step in a study is to find answer to the question of how many cases should be studied with in the planning, design and application phases. In this study, practical sample size tables have been formed in the rates of three different type I error and three different tests and different disagreement rates. Through these tables, it would be possible to study with suitable and enough number of cases for the design of the study and the outcome variable. Besides, in this study, the agreement statistics used in the calculation of the agreement between the raters/methods in clinical studies and the superiorities of these agreement statistics to each other and the bias of these statistics and the mistakes made have been focused on.

#### *Limitations*

In this study, only one measurement of two raters belonging to the agreement statistics commonly used in literature; and sample size tables for the agreement statistics between these measurements have been formed.

Table V: The necessary sample size for the agreement coefficients given developed for the agreement coefficients by Kilem Gwet

| $(\pi_A-\pi_E)$ | Relative error (EB) | N=30 | N=100 | N=500 | N=1000 | N=10000 |
|---|---|---|---|---|---|---|
| **0.10** | 0.10 | 30 | 99 | 477 | 909 | 5000 |
| | 0.20 | 30 | 96 | 417 | 714 | 2000 |
| | 0.30 | 29 | 92 | 345 | 526 | 1000 |
| | 0.40 | 29 | 86 | 278 | 385 | 588 |
| | 0.50 | 28 | 80 | 222 | 286 | 385 |
| **0.20** | 0.10 | 30 | 96 | 417 | 714 | 2000 |
| | 0.20 | 29 | 86 | 278 | 385 | 588 |
| | 0.30 | 27 | 74 | 179 | 217 | 270 |
| | 0.40 | 25 | 61 | 119 | 135 | 154 |
| | 0.50 | 23 | 50 | 83 | 91 | 99 |
| **0.30** | 0.10 | 29 | 92 | 345 | 526 | 1000 |
| | 0.20 | 27 | 74 | 179 | 217 | 270 |
| | 0.30 | 24 | 55 | 99 | 110 | 122 |
| | 0.40 | 21 | 41 | 61 | 65 | 69 |
| | 0.50 | 18 | 31 | 41 | 43 | 44 |
| **0.40** | 0.10 | 28 | 86 | 278 | 385 | 588 |
| | 0.20 | 25 | 61 | 119 | 135 | 154 |
| | 0.30 | 21 | 41 | 61 | 65 | 69 |
| | 0.40 | 17 | 28 | 36 | 38 | 39 |
| | 0.50 | 14 | 20 | 24 | 24 | 25 |
| **0.50** | 0.10 | 28 | 80 | 222 | 286 | 385 |
| | 0.20 | 23 | 50 | 83 | 91 | 99 |
| | 0.30 | 18 | 31 | 41 | 43 | 44 |
| | 0.40 | 14 | 20 | 24 | 24 | 25 |
| | 0.50 | 10 | 14 | 16 | 16 | 16 |
| **0.60** | 0.10 | 27 | 74 | 179 | 217 | 270 |
| | 0.20 | 21 | 41 | 61 | 65 | 69 |
| | 0.30 | 15 | 24 | 29 | 30 | 31 |
| | 0.40 | 11 | 15 | 17 | 17 | 17 |
| | 0.50 | 8 | 10 | 11 | 11 | 12 |
| **0.70** | 0.10 | 26 | 67 | 145 | 169 | 200 |
| | 0.20 | 19 | 34 | 46 | 49 | 51 |
| | 0.30 | 13 | 18 | 22 | 22 | 23 |
| | 0.40 | 9 | 11 | 12 | 13 | 13 |
| | 0.50 | 6 | 8 | 8 | 8 | 8 |
| **0.80** | 0.10 | 25 | 61 | 119 | 135 | 154 |
| | 0.20 | 17 | 28 | 36 | 38 | 39 |
| | 0.30 | 11 | 15 | 17 | 17 | 17 |
| | 0.40 | 7 | 9 | 10 | 10 | 10 |
| | 0.50 | 5 | 6 | 6 | 6 | 6 |
| **0.90** | 0.10 | 24 | 55 | 99 | 110 | 122 |
| | 0.20 | 15 | 24 | 29 | 30 | 31 |
| | 0.30 | 9 | 12 | 13 | 14 | 14 |
| | 0.40 | 6 | 7 | 8 | 8 | 8 |
| | 0.50 | 4 | 5 | 5 | 5 | 5 |
| **1.00** | 0.10 | 23 | 50 | 83 | 91 | 99 |
| | 0.20 | 14 | 20 | 24 | 24 | 25 |
| | 0.30 | 8 | 10 | 11 | 11 | 11 |
| | 0.40 | 5 | 6 | 6 | 6 | 6 |
| | 0.50 | 4 | 4 | 4 | 4 | 4 |

**$(\pi_A-\pi_E)$: The difference between of the overall agreement probability and the chance-agreement probability; Relative error (EB); Effect size**

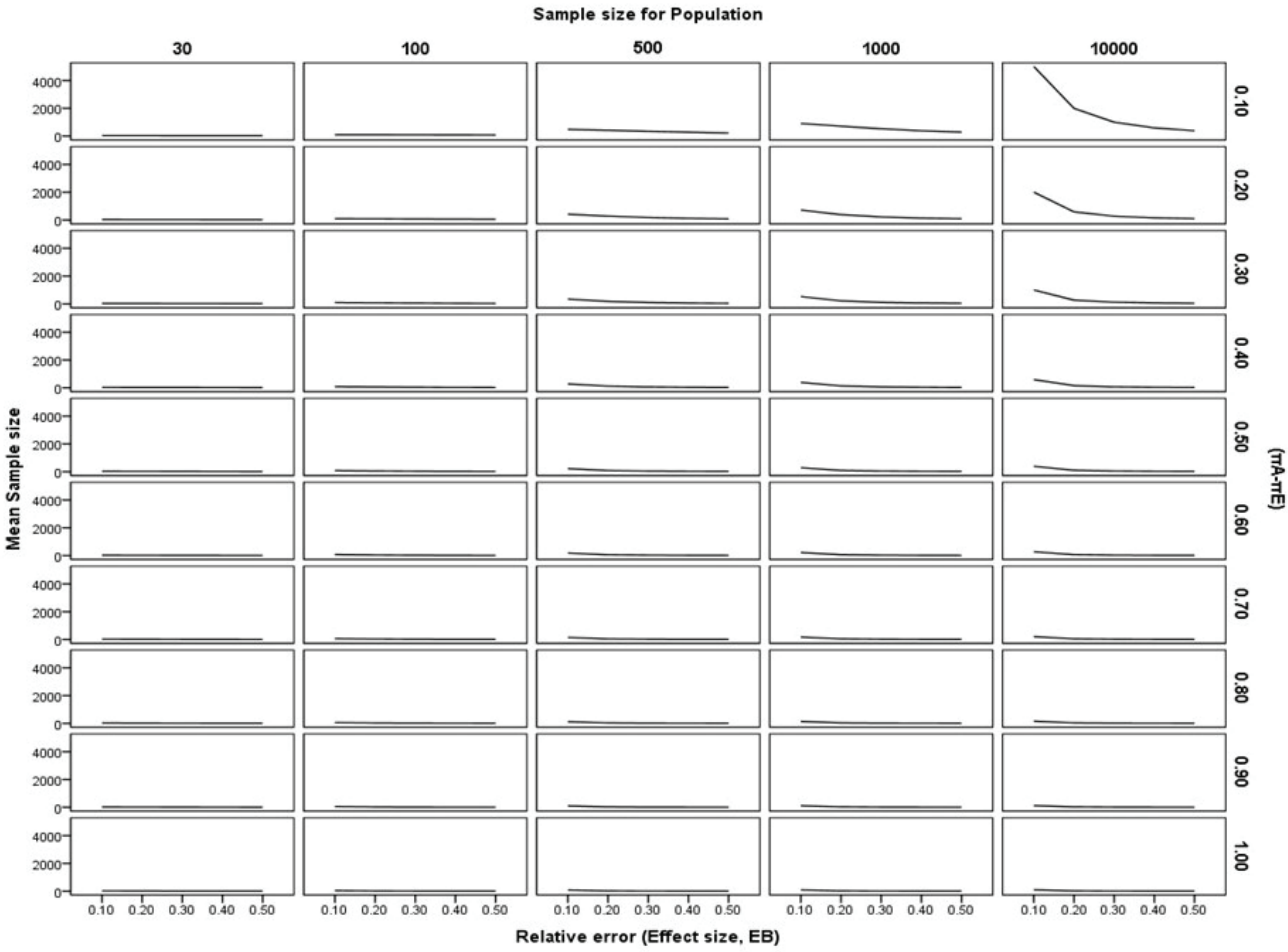


Figure 4: The necessary sample size for the agreement coefficients given developed for the agreement coefficients by Kilem Gwet

## Discussion

Determination of sample size is one of the most important and even the most difficult steps in the planning of clinical studies. Studying with enough samples is quite important because of scientific, economical and ethical reasons. Studying on a sample large enough is the most important factor guaranteeing the findings to be obtained from the research, its scientific validity and reliability. While studying with less than enough samples in number in scientific studies would decrease the power of the study, studying with more than enough samples in number would lead to a futile effort and resource waste. Besides, in the determination of sample size with power analysis in clinical studies, type I error, power and effect size have to be known. The importance value's being small and power's being large are the reasons increasing the sample size. Besides, it is necessary to study with more samples in small effect sizes [2,3].

One of the most common mistakes made in the agreement studies between the raters is the confusion of the concepts of agreement and relationship. When our outcome variable is in a continuous state, instead of getting help from ICC, Pearson or Spearman correlation coefficients giving the relationship between the two continuous variables are used to test the agreement between the raters. Generally, when the measurement agreement is high, it is possible to obtain information that the agreement with these two tests is also good. However, it is a mistake to use these two tests in the use of measurement agreement analysis [11].

In the case when our outcome variable is in a categorical structure, instead of Kappa statistic value, Mc-Nemar test used in the testing of whether there is a difference between the results of the two raters is preferred to test the agreement. However, it is a mistake to use this test in agreement analysis, too. Besides, it is also suggested by the researchers

that Kappa statistic is affected from the prevalence and that very intensive importance have to be given when it is used in agreement studies as agreement coefficient [8]. Thus, while calculating the sample size, the design of the study has to be known very well, too.

As a result, with the accurate determination of enough minimum sample size suitable for the design of a study and the state of the result variable at the beginning of the test, besides providing the reliability of the study results, the waste of samples would also be avoided.

## References

1. Kanık EA, Erdoğan S. Değerlendiriciler arası uyumun saptanması. Mersin Univ Tıp Fak Derg 2004; 5: 430-7.
2. Özdamar K. . Modern bilimsel araştırma yöntemleri. Eskişehir: Kaan Kitabevi, 2003.
3. Süt N. Klinik araştırmalarda örneklem sayısının belirlenmesi ve güç (power) analizi. RAED Dergisi 2011; 3: 29-33.doi: 10.2399/raed.11.005
4. Gwet K. Kappa Statistics is not satisfactory for assessing the extent of agreement between raters. Series: Statistical Methods Inter-Rater Reliability Assessment 2002; 1: 1-5.
5. Machin D, Tan S B, Champbell MJ, (editors). Sample size tables for clinical studies. Singapore: BMC Books, 2009
6. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Mathem Stat Psychol 2008; 61: 29-48. doi:10.1348/000711006X126600
7. Viera AJ. Garrett JM. Understanding interobserver agreement: the kappa statistics. Fam Med 2005; 37: 360-3.
8. Kanık EA, Erdoğan S, Orekici Temel G. İki sonuçlu tanı testlerinde iki hekim arasındaki uyum istatistiklerinin prevelanstan etkilenme durumları. İnönü Üniversitesi Tıp Fakültesi Dergisi 2012; 19: 153-8. doi: 10.7247/jiumf.19.3.5
9. Fleiss JL, Shrout PE. Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin 1979; 86:420-8.
10. Alpar R. Spor, sağlık ve eğitim bilimlerinden uygulamalı istatistik ve geçerlik-güvenirlik. Ankara: Detay Yayıncılık, 2012.
11. Erdoğan S, Kanık EA. Rasgele ve sistematik hataların sınıf içi ve uyum korelasyon katsayıları ve bland ve altman yöntemi üzerine etkileri: bir simülasyon çalışması. VIII. Ulusal Biyoistatistik Kongresi; 20-22 Eylül 2005; Bursa. Bursa: Uludağ Üniversitesi; 2005.
12. Bonnett DG. Sample size requirements for estimating intraclasss correlations with desired precision. Stat Med 2002; 21: 1331-5. doi: 10.1002/sim.1108
13. Haley DT, Thomas P, Petre M, Roeck AD. Using a new inter-rater reliability statistics. Technl Rep 2008; 15: 14-23.
14. Kanık EA, Orekici Temel G, Erdoğan S, Ersöz Kaya I. Comparison of agreement statistics in case of multiple-raters and diagnostic test being categorical: a simulation study. J Turgut Ozal Med Cent 2012;19: 220-7. doi:10.7247/jtomc.19.4.4
15. Gwet KL. Handbook of inter-rater reliability. USA: Advanced Analytics, LLC, 2014.
16. Gwet KL. Variance estimation of nominal-scale inter-rater reliability with random selection of raters. Pyschometrica 2008;73:407-30. doi: 10.1007/S11336-007-9054-8