Seminars in Oncology Nursing 39 (2023) 151398

Contents lists available at ScienceDirect

Seminars in Oncology Nursing

journal homepage: https://www.journals.elsevier.com/seminars-in-oncology-nursing

An Overview of the Fundamentals of Data Management, Analysis, and Interpretation in Quantitative Research

Grigorios Kotronoulas[a,*], Susana Miguel[b], Maura Dowling[c], Paz Fernández-Ortega[d], Sara Colomer-Lahiguera[e], Gülcan Bağçivan[f], Eva Pape[g], Amanda Drury[h], Cherith Semple[i], Karin B. Dieperink[j], Constantina Papadopoulou[k]

[a] *Reader, School of Medicine, Dentistry & Nursing, University of Glasgow, Glasgow, Scotland, UK*
[b] *Clinical Nurse Specialist, Department of Head and Neck and ENT Cancer Surgery of the Portuguese Institute of Oncology of Francisco Gentil, Lisbon, Portugal*
[c] *Senior Lecturer, School of Nursing and Midwifery, University of Galway, Galway, Ireland*
[d] *Associate Professor, Catalan Institute of Oncology and Faculty of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain*
[e] *Senior Nurse Scientist, Institute of Higher Education and Research in Healthcare (IUFRS), Faculty of Biology and Medicine, University of Lausanne, and Lausanne University Hospital, Lausanne, Switzerland*
[f] *Associate Professor, School of Nursing, Koc University, Istanbul, Turkey*
[g] *Clinical Nurse Specialist, Department of Gastrointestinal Surgery, Cancer Center, Ghent University Hospital, Ghent, Belgium*
[h] *Associate Professor, School of Nursing, Psychotherapy and Community Health, Dublin City University, Dublin, Ireland*
[i] *Reader, School of Nursing, Institute of Nursing and Health Research, Ulster University, Belfast, UK*
[j] *Professor, Department of Clinical Research, University of Southern Denmark, Department of Oncology, Odense University Hospital, Odense, Denmark*
[k] *Reader, School of Health and Life Sciences, University of the West of Scotland, South Lanarkshire, Scotland, UK*

ARTICLE INFO

Key Words:
Quantitative studies
Data analysis
Data management
Interpretation
Empirical research
Statistics

ABSTRACT

Objectives: To provide an overview of three consecutive stages involved in the processing of quantitative research data (ie, data management, analysis, and interpretation) with the aid of practical examples to foster enhanced understanding.
Data Sources: Published scientific articles, research textbooks, and expert advice were used.
Conclusion: Typically, a considerable amount of numerical research data is collected that require analysis. On entry into a data set, data must be carefully checked for errors and missing values, and then variables must be defined and coded as part of data management. Quantitative data analysis involves the use of statistics. Descriptive statistics help summarize the variables in a data set to show what is typical for a sample. Measures of central tendency (ie, mean, median, mode), measures of spread (standard deviation), and parameter estimation measures (confidence intervals) may be calculated. Inferential statistics aid in testing hypotheses about whether or not a hypothesized effect, relationship, or difference is likely true. Inferential statistical tests produce a value for probability, the *P* value. The *P* value informs about whether an effect, relationship, or difference might exist in reality. Crucially, it must be accompanied by a measure of magnitude (effect size) to help interpret how small or large this effect, relationship, or difference is. Effect sizes provide key information for clinical decision-making in health care.
Implications for Nursing Practice: Developing capacity in the management, analysis, and interpretation of quantitative research data can have a multifaceted impact in enhancing nurses' confidence in understanding, evaluating, and applying quantitative evidence in cancer nursing practice.

Introduction

Quantitative research assumes that the constructs under study can be measured. As such, quantitative research aims to process numerical data (or numbers) to identify trends and relationships and to verify the measurements made to answer questions like who, how much, what, where, when, how many, and how.[1,2] In this context, the processing of numerical data is a series of steps taken to help researchers and consumers of research (eg, health professionals, patients, policy makers, and the public) make meaning from the data. The process itself can involve a lot of negotiation, mainly because

* Address correspondence to: Grigorios Kotronoulas, School of Medicine, Dentistry & Nursing, University of Glasgow, Glasgow, Scotland.
E-mail address: Grigorios.kotronoulas@glasgow.ac.uk (G. Kotronoulas).

https://doi.org/10.1016/j.soncn.2023.151398

what once were random numbers start to become more concrete, yet its meaning must be explored and explained carefully to establish the degree to which the evidence answers the research question(s).

This article aims to provide an overview of three consecutive stages involved in the processing of quantitative research data (ie, data management, analysis, and interpretation) with the aid of practical examples to foster enhanced understanding.

Data Management

Typically, a considerable amount of numerical data is collected that require analysis. Also, typically, the data can be disorganized and made up of separate bits of information. Imagine that a group of researchers have collected a set of numerical data as part of a quantitative study. At a very early stage, the numbers will probably look like the screenshot in Fig 1. What information do they convey? When numbers look this way, their meaning is unclear.

Crucially, the starting point is to carefully prepare a data set ready for analysis. Numbers (also known as raw data) must be put in a form that makes them suitable for analysis. This is called data management; its purpose is to make the data analyzable. At the core of quantitative data management is the construction and definition of variables. A variable is defined as anything that can be measured that varies. In health care, variables may represent things that vary from one person to the next (eg, country of origin, type of cancer) and even within the same person (eg, temperature, neutrophil counts). Practically, variables will contain the quantitative data to be statistically analyzed. Much of what happens to variables during data analysis depends on their type. One fundamental classification is based on the values and units of measurement attached to variables. The algorithm in Fig 2 can quickly help you identify a variable as dichotomous, categorical, or metric.

Dichotomous variables only have two distinct values involved; this is why they are also known as binary. Categorical variables are variables that have three or more values; they can be either nominal or ordinal. In nominal variables the order of the values does not matter. For instance, the variable country of origin might comprise a list of countries; however, which country goes first, which second, and so on does not matter. If the order does matter and there is no fixed unit of measurement, then the variable becomes ordinal. Consider here a numerical scale from 0 to 10 that indicates the level of pain. Pain graded as 8 will always mean more severe pain compared to a value of 3; there is thus an ascending order of pain severity that is fixed from 0 to 10. Pain severity has no attached unit of measurement on this scale. Where the order of the values matters and there is a fixed unit of measurement involved, then we talk about metric variables. Weight, blood pressure, and time are all metric variables because the order of values is important, and there is a fixed unit of measurement attached to these values. Blood pressure might be measured in mm Hg, and a value of 130 mm Hg will always mean higher systolic blood pressure than 120 mm Hg.

Codes must also be attached to quantitative variables as necessary to help with the interpretation of results. This is particularly true for dichotomous and categorical variables. For instance, as part of coding, values of 1 and 2 in the variable gender would be assigned a man and woman descriptor, respectively. In many cases, a code book is created. Subsequently, the data will be entered in a file (Fig 1), using a data processing software (eg, Microsoft Excel) and "cleaned." Data are thoroughly checked for inconsistencies or errors in data entry (eg, due to mistyping) or for missing values (see Fig 1). Missing values can be assigned a code in the data set (eg, 999) for ease of interpretation. The goal is to minimize the risk for inconsistencies, errors, and missing values to have a major impact on the final results. Data cleaning, as an essential aspect of quality assurance and a determinant of validity, should not be an exception. In quantitative study protocols, it is advised inclusion of a data-cleaning plan.[4]

Running descriptive statistics (see relevant section further below) can help spot most errors. Some of the most common errors include:

- Inconsistent data entry or misspelling. For example, data for gender might be entered as "F," "f," "fem," "female," or "1." These can cause problem with coding and interpretation. Frequency tables allow auditing all the text that was typed in originally.
- Out-of-range values. For example, a respondent's pain score on a 0-10 visual analogue scale might have been mistyped as 13 instead of 3. Without correction of this error, data analysis could lead to inflated pain scores for the sample and misleading conclusions. A frequency table would again be valuable in identifying out-of-range values.

Relatedly, any missing values will also require special consideration. Missing values simply mean that for one or more variables or for a number of study participants data are not available. It may well be because study participants skipped a question or questionnaire, missed a measurement point, or because they dropped out of the study completely. A large number of missing values creates problems with the analysis because it leads to an imbalance in the data set, which might interfere with the validity of the data and the accuracy of the conclusions drawn from the analysis. Although concrete benchmarks regarding what percentage of missing values is

Participant ID	Age (years)	Family status	Group (A or B)	Distant Mx	CA-125 (U/mL)	QoL score
1	9	1	A	Yes	36	33
2	76	2	C	No	66	89
3	55	3	B	Yes	189	67
4		1	A	Yes	22	10
5	34	1	B	No	61	
6	16		B		77	50
7	43		B	Yes	46	88
8	34	1	B	No	33	100
9	78	1	A	No	165	12
10	68	1	A	No		24
11	92	3	A	Yes	19	49
12	76	2	A	Yes	23	32
13	62	2	B	No	122	990

FIG 1. Snapshot of a data set with raw quantitative data collected as part of a fictitious study among patients with ovarian cancer receiving one of two treatments (A or B). Shaded cells indicate missing values. Values highlighted in red indicate possible errors in data entry.

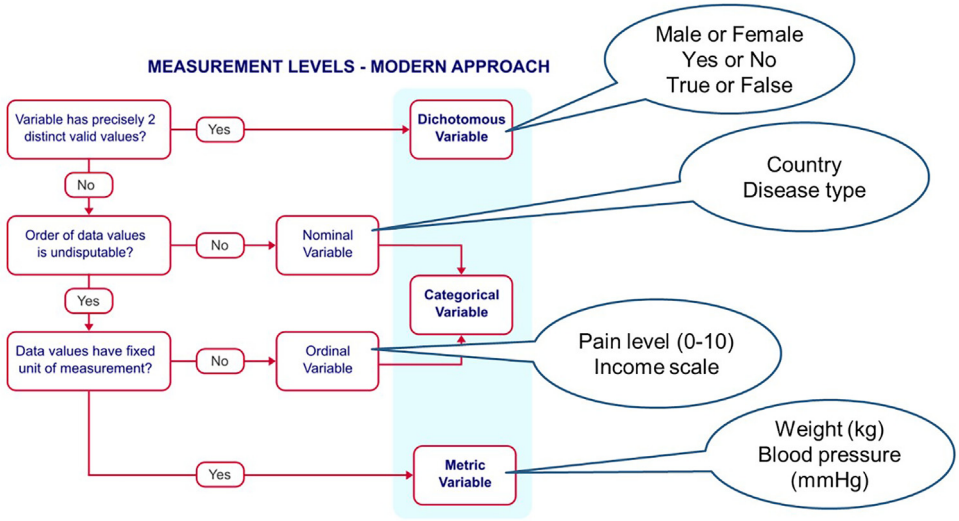

FIG 2. Modern approach to the classification of variables. Adapted from van den Berg.[3]

acceptable do not exist, there are several techniques to help analysts deal with missing values in their data set. The commonly used techniques to handle missing values include deletion of the missing data, substitution, and imputation; for additional information, please see examples by Kang.[5] Although such techniques have their merit, it is advisable that careful consideration is given on how to prevent missing values during the data collection stage. In any case, having attended to the above actions, a data set is created, ready for analysis.

Data Analysis

Quantitative data analysis involves the use of statistics. Statistics will always analyze variables to help you make sense of numerical data drawn from a sample. This is where the raw data (numbers) become results or evidence. In most cases (if not all), data analysis begins with devising a clear analysis plan (or statistical analysis plan [SAP]) to ensure that statistics are aligned to the analyses required to help address the research questions that have been set in a given study.

Statistics have two functions. First, statistics can be descriptive. Descriptive statistics generate summaries of the variables in a data set to show what is typical for the sample. Second, statistics can be inferential. Inferential statistics aid in exploring links between variables and making inferences. This means that depending on the nature of the research, statistics can be used to show whether a new treatment is effective, to investigate whether two or more variables might be related to one another, or to reveal how similar or different two samples might be. Importantly, inferential statistics also indicate whether an observed effect, relationship, or difference is a chance finding or it is likely to be true and existing in reality.

Please note that statistics never *prove* anything. They only help to quantify exactly how certain or uncertain effects, relationships, similarities, or differences are—although one can never be 100% sure. The whole approach is probabilistic.[2] Imagine it as the middle ground or a gray area between black and white. One can never say for sure that something *will* happen to everyone or at all times; something is only likely or unlikely to happen; thus, there is a probability attached to it. If the odds are high that something can happen, then there is greater certainty (but never 100% certainty) that this may actually happen in the real world for the majority of patients. For example, if statistics show that one in five patients with head and neck cancer will develop oral mucositis due to radiation to the oral cavity,[6] then the typical patient from this target population has a 20% risk to experience oral mucositis. However, this probability may increase or decrease depending on individual characteristics that put a person (patient) at greater or lower risk of oral mucositis. Therefore, there is always going to be some degree of uncertainty in any given statistic.

Descriptive Statistics

Descriptive statistics summarize the data to describe how the sample looks like. At its simplest, this information can be reported as frequencies (ie, total numbers and percentages). Frequencies only apply to dichotomous and categorical variables. In published articles, you will find frequencies given as text or displayed in tables or graphs (Fig 3). The idea of tables and graphs is to condense the information and present it in a way that is visually attractive and easily comprehensible. Importantly, tables and graphs must be self-explanatory as stand-alone sources of information, including details in the heading, descriptor(s), and footnote(s) to allow the reader to fully understand the summarized data being presented.

Descriptive information can also be reported using special measures that indicate one of the following:

(a) The central position of the data. These are known as measures of central tendency; metric variables can be analyzed this way.
(b) How spread out the data are. These are known as measures of spread or dispersion; again, metric variables can be analyzed this way.
(c) What the data might look like in the actual population. These are called parameter estimation measures; they apply to all types of variables.

Measures of Central Tendency

Measures of central tendency indicate the central position of the data in the data set. These are helpful measures that can quickly show you how the data tend to cluster around a middle value. The arithmetic mean (nonscientifically known as the average) probably is the measure that you are most familiar with. It is the sum of a set of numbers divided by the count of numbers in the set (eg, 3 + 5 + 7 + 7 + 8 = 30 / 5 = 6). Another common measure of central tendency is the median. The median is the middle in a sorted, ascending or descending, list of numbers (eg, 3, 5, 7, 7, 8 = 7). The mode (ie, the most frequent number in a set of data values, such as 3, 5, 7, 7, 8 = 7) is probably the measure least commonly used today to describe metric variables. However, the mode can be used to summarize

Table 1. Baseline demographic and clinical patient characteristics (*n*=20)

		Frequency	%
Age	Mean (SD)	67.1 (8.62)	
	Median	69.5	
	Range	32 (51-83)	
	IQR	13	
Time since diagnosis (days)[a]	Mean (SD)	284.6 (404.3)	
	Median	118.0	
	Range	1401 (20-1421)	
	IQR	267	
Gender	Male	13	65.0
Educational background	High school	18	90.0
	Some college	1	5.0
	University	1	5.0
Marital status	Married/partnered	13	65.0
	Single	2	10.0
	Divorced	3	15.0
	Widowed	2	10.0
Employment status	Employed	2	10.0
	Unemployed	4	20.0
	Retired	14	70.0
Type of disease	NSCLC	16	80.0
	SCLC	2	10.0
	Other (e.g. mesothelioma)	2	10.0
Disease stage[b]	Local	9	47.4
	Metastatic	10	52.6
Co-morbid illnesses	Yes	5	25.0

[a]n=11
[b]n=19
Abbreviations: IQR – Interquartile range; NSCLC – non-small cell lung cancer; SCLC – small cell lung cancer

FIG 3. Example table providing a detailed overview of the sample's demographic and clinical characteristics, using descriptive statistics. Adapted from Kotronoulas et al.[7]

categorical variables. The meaning of these measures becomes more obvious when you compare two or more groups, for example, when you compare the mean weight (or median weight) between two geriatric cancer groups, a prefrail and a frail one (Fig 4).

Measures of Dispersion

Knowing the central point in a data set is quite useful; however, it can only be meaningful if you also know how spread out the data are. This gives us an indication of how diverse the study sample is on a particular variable. See the two data sets A and B in Fig 4; data set A has been created with data from a prefrail geriatric cancer patient group; data set B comes from a frail geriatric cancer patient group. Both data sets refer to the same variable: body weight. The data sets have the same median (70 kg) and roughly the same mean; in other words, the center of these two distributions is practically the same. However, the data sets are quite different in terms of how dispersed the data are around the mean (or median). In data set A, data are quite spread out; see how wide the line above the data is. Conversely, in data set B, the data seem to be a lot closer together. This means that the patients in that group (sample) were more similar or that the sample was more homogenous. Here, we have two measures of spread as they are called. The range is easy to calculate as it is the distance between the smallest (minimum) and the largest point (maximum) in a distribution of data. If we were to compare the two data sets, see how data set A has a range of 60, whereas data set B has a range of 18. This immediately gives you a basic idea about the spread of the data.

An even better measure of spread is the standard deviation (often reported as SD). The standard deviation is the mean (average) distance between each data point and their mean. A low standard deviation indicates that the values tend to be close to the mean of the set, whereas a high standard deviation indicates that the values are spread out over a wider range. Again, you see in Fig 4 that with a standard deviation of 6 (vs 25 in data set A), data set B seems to be more homogenous as the values seem to cluster together. Knowing the standard deviation has implications when inferential statistics are used (more on this to follow).

Now, data in a data set can be spread in all sorts of ways. In some cases, most data will be on the left or most on the right or in no particular direction. In other cases, the central point may be in the center of the distribution; this is called a normal distribution. The normal distribution has a bell-like shape and is symmetrical, meaning that 50% of the data will be on the left of the center and the other half on the right. In a normal distribution, the mean and the median will take the same value; the mean becomes the middle point of the distribution. Most variables in a population follow the normal distribution. This might not always be the case in a sample, whereby the influence of sample size or sampling method might be translated into inclusion of participants with only certain characteristics or experiences, which can polarize the data collected. Data for some variables may look plausibly normally distributed; however, for other variables data may look more skewed to the right (ie, the higher data values are fewer; positively skewed distribution) or to the left (ie, the lower data values are fewer; negatively skewed distribution). When measures of central tendency are to be calculated, it is always a good idea to check the distribution of the data. Where the data seem plausibly normally distributed, either the mean or the median can be calculated to provide similar information. Where the data are skewed to

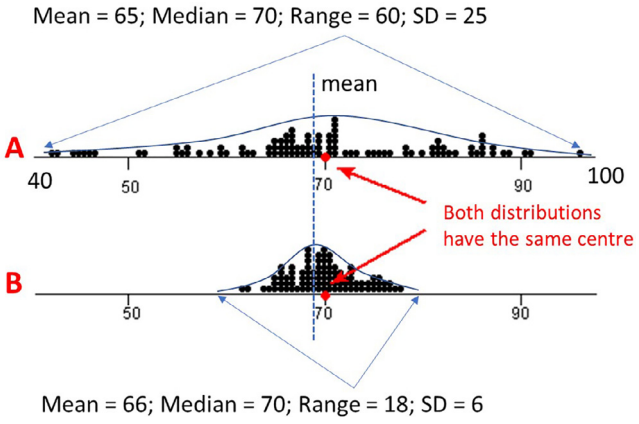

FIG 4. Example of measures of central tendency and dispersion related to two independent data sets (A and B) that relate to the same variable (weight).

Table 2: Baseline characteristics of the study participants

Variable	Control	Triamcinolone	P
Age	56.46±9.36	58.53±8.89	0.384*
Sex			
Male	17 (56.7)	18 (60)	0.793#
Female	13 (43.3)	12 (40)	
Smoking			
Yes	11 (36.7)	14 (46.7)	0.601#
Status of denture			
Yes	21 (70)	19 (63.3)	0.785#

Values are frequency (%) for categorical and mean±SD for quantitative variables.
*Resulted from independent *t*-test; #Resulted from Chi-square test. SD=Standard deviation

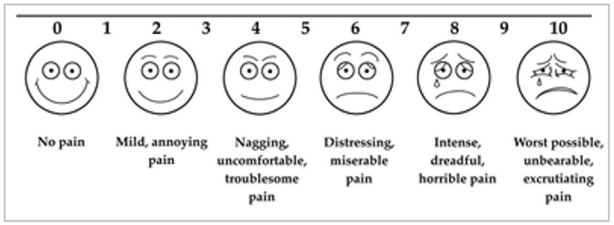

Table 3: Comparison of the mean pain score in the study groups

	Control	Triamcinolone	P*
Pain score at week 0	5.34±0.78	5.36±1.29	0.935
Pain score at week 1	5.16±0.68	4.83±1.91	0.378
Pain score at week 2	4.96±0.67	4.56±1.67	0.428
Pain score at week 3	4.84±0.92	2.80±2.15	<0.001
Pain score at week 4	4.69±0.77	2.20±2.02	<0.001

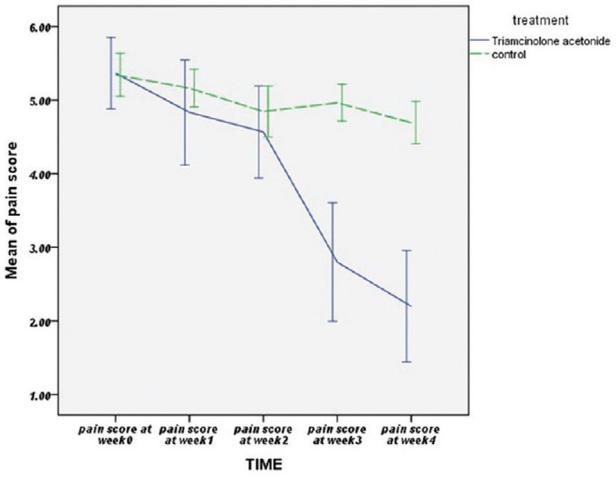

FIG 5. Measures of central tendency and dispersion reported in a published article. Adapted from Pakravan et al.[8]

the right or the left, then the median is a better option. This is because the mean is easily influenced by too small or too large values in the data set and, as such, can provide an artificially inflated or deflated summary of the variable under consideration. Conversely, the median is unaffected by extreme values.

A data distribution also gives an idea of where, percentage wise, a certain value falls by making use of the standard deviation. There is an empirical rule associated to the normal distribution that can help you better understand this. In a normal distribution, you can expect that 68% of all values will fall within 1 standard deviation to the left or to the right of the center; 95% of the values will fall within two standard deviations, while almost all values will fall within three standard deviations (98%). Let us examine a practical example in Fig 5. Pakravan et al[8] tested the use of a triamcinolone patch compared to placebo as an intervention to treat oral mucositis induced by radiotherapy for patients with head-and-neck cancer. Fig 5 shows descriptive statistics for the two groups. Please note how variables like sex, smoking, and status of denture are summarized using total numbers and percentages. Age is presented as the mean and standard deviation. See how patients in the triamcinolone group were on average older than the control (58.5 vs 56.5 years). The standard deviations were comparable (8.9 vs 9.4), perhaps the control group was just a bit more heterogenous as the slightly higher standard deviation implies. Knowing the standard deviation, you also know that 68% of the data about age will fall within one standard deviation on the left or right of the mean, or for example, that 68% of data in the triamcinolone group will be between roughly 49.6 and 67.4 years of age.

Pakravan et al[8] also measured patients' pain on a 0-10 numerical scale, and they did so at baseline, the week before the trial started (that is week 0), and after the treatment was given for 4 consecutive weeks. They have provided descriptive statistics in the form of means and standard deviations for each weekly measurement and each group (see Fig 5). All these numbers may be hard to grasp. This is why the article also includes a helpful graph where you can see the change in mean scores over time. The blue line is the treatment group. Each time point is the mean pain score. Patients in the treatment group seem to report lower scores on average compared with the control group; see the steep decline to the blue line over time. In addition, at each point you see a vertical line, which indicates the standard deviation (ie, the spread of the data around the mean). See how the vertical lines change in length; this is because standard deviation changes at each independent measurement over time. From the graph, you can quickly tell that pain scores for the control group where much closer together (the lines are shorter, meaning that the spread was lower), whereas for the intervention group the lines are longer because the standard deviations were higher and the data more spread out, meaning that the intervention sample was less homogenous. In our example, this reduced homogeneity means that the reported pain scores were quite variable after the intervention was given to the study participants.

Parameter Estimation Measures (Confidence Intervals)

A confidence interval (CI) simply is a way to measure how well a sample represents the wider population that is being studied. This is important in evidence-based practice. Clinicians need to have an estimate of a population parameter that comes from investigating just one sample from this same population. Clinicians also need to know how accurate this estimate is; in other words, how likely the sample is to accurately reflect the wider population. Suppose that the parameter of interest is the arithmetic mean of posttreatment survival gain in months in the wider population of patients with cancer. A sample of patients with cancer is studied, and the mean survival gain comes back as 2.5 months. What is our confidence that this value (which comes from just one sample) accurately reflects mean survival gain in the wider cancer patient population?

The mean survival gain in the wider cancer patient population is unknown; however, the mean from the sample can help you make an estimation. This is the function of calculating a confidence interval.[9,10] The confidence interval is a range of values. This range indicates how likely it is for the values to include the true value of a population parameter (such as the arithmetic mean) with a certain degree of confidence.[2] A confidence interval is often expressed as a percentage. The 95% confidence interval is most commonly reported in published articles. In our example, the 95% confidence interval can

be calculated as -0.01 to 5.01 months.[7] In the hypothetical scenario that we were to run 100 independent, identical studies involving 100 different cancer patient samples and computed a 95% confidence interval for each sample, then exactly 95 of the 100 confidence intervals would contain the true mean survival gain, and exactly 5 of the 100 confidence intervals would not.[11]

Consider another practical example here. If a researcher studied weight management in 100 frail geriatric patients with cancer, they might have found that the sample had a mean weight of 73 kg with a standard deviation of 17. Clinicians would ask: How confident can we be that this mean weight reflects the wider population of frail geriatric patients with cancer? A 95% confidence interval can be calculated to show that if, in the hypothetical case that 100 similar samples of geriatric patients with cancer were investigated and 100 confidence intervals were computed, then 95 of them would contain the true population mean weight and that would be between 69.7 and 76.3 kg. Many journals today require researchers to report confidence intervals to increase the meaningfulness of the information for clinicians.

Inferential Statistics

This branch of statistics aims to test hypotheses to return a probability about whether or not a hypothesized effect, relationship, or difference is likely true. In inferential statistics, you have two hypotheses.[12] The null hypothesis (H0) states that there is no effect, relationship, or difference. For instance, the null hypothesis might state that severe pain IS NOT related to frequent nighttime awakenings in patients with advanced cancer. The alternative hypothesis (H1) states the exact opposite; this is actually what the researchers are after, concrete evidence to allow them to say that severe pain IS linked (ie, increases the possibility) for patients with advanced cancer to spend long hours awake at night.

The null hypothesis can either be true or false. In effect, the goal is to reject the null hypothesis when the null is false because this supports the alternative hypothesis as being true. In that sense, hypothesis testing involves making decisions about when to reject or not reject the null hypothesis. The only way to know this is through analyzing data from a sample via use of inferential statistics.[12]

Researchers always want to be able to correctly decide to reject a null hypothesis when it is actually false. Similarly, they want to correctly decide to not reject a null hypothesis when it is true (Fig 6). Of course, this is not always easy. Therefore, researchers set criteria about how confident they wish to be in their decision-making. These criteria come in the form of probabilities. Usually, researchers only allow 20% or less chance for them to fail to reject a null hypothesis that is false, for instance, to fail to reject a null hypothesis that says that a new treatment does not work when it actually works. This probability is called beta or Type II error. The opposite (ie, 1 - beta) is usually called the statistical power of the study, and it is set at 80% or above (100% - 20% = 80%). Also, researchers usually only allow 5% or less chance to reject a null hypothesis that is true (ie, to say that a treatment works although in reality the treatment is ineffective). This probability is called Type I error or alpha or the significance level alpha.

To decide whether to reject or not reject a null hypothesis, researchers use statistical tests. The test takes into account the distribution of the data and the type of variables (eg, metric) involved in the hypothesis. The test uses the numbers attached to the variables to produce a value for the probability that the null hypothesis is true. There are many statistical tests; the basic ones are:

- Chi-square – compares two dichotomous variables.
- Pearson's r coefficient – shows how two metric variables correlate linearly.
- Student's *t* test – compares the means between two independent groups on a metric variable.
- Analysis of variance – compares the means of three or more independent groups.
- Regression analysis – shows the effects of one variable on another variable when every third variable stays the same.

Shreffler and Huecker[13] offer a nice overview of statistical tests as they apply to research questions and variables.

For statistical tests to be used appropriately for data analysis, the data must meet certain assumptions. For example, continuous data on a metric variable must be normally distributed for a Student's *t* test to be conducted. Tests that assume that the data from the sample are normally distributed are called *parametric*. In fact, assessing the normality of data in metric variables is a prerequisite for many statistical tests and, therefore, should be actioned early on to allow for selection of the most appropriate test. Data in metric variables can be assessed for normality either visually (eg, by examining their distribution on a histogram) or statistically.[14] If the data do not meet this assumption, it is necessary to consider alternatives. This is usually a decision between two options:

a. Keep the variable unchanged and use an equivalent, *nonparametric* test. Nonparametric tests are also known as distribution free tests exactly because they do not assume anything about how the

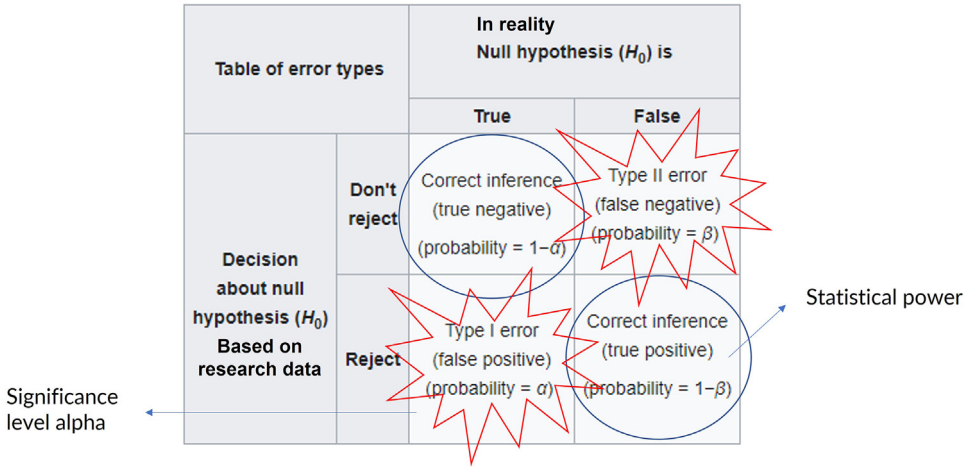

FIG 6. Criteria for decision-making in hypothesis testing.

data may be distributed. In this example, the equivalent to a Student's *t* test would be the Mann-Whitney test.

b. Convert the metric variable into a categorical variable to meet the assumptions of a test that is suitable for categorical variables (eg, the Chi-square test). For instance, age (metric variable) could be converted into age groups (categorical variable). Alternatively, a questionnaire score for quality of life could be converted to a high and low score.

Deciding how to categorize data for abstract metric variables such as quality of life may be difficult because no specific benchmarks exist. A decision may be to use a standardized grouping with equal scores for each group. Consider total scores on the Functional Assessment for Cancer Therapy-Colorectal (FACT-C); this is a questionnaire that measures quality of life in patients with colorectal cancer. Total FACT-C scores with a possible range of 0-136 may be arbitrarily broken down to low (0-45), moderate (46-90), and high (91-136) quality of life. As such, nominal categorical variable is created from an original metric one. However, it is possible that some groups (eg, low) might end up having too few participants to allow chi-square analysis; adequate numbers within each subgroup is one of the assumptions of chi-square tests. An alternative approach may be to dichotomize the original FACT-C variable, so that half the sample fall into a "higher" and half into a "lower" quality of life group. If taking this approach, the median score for the sample would act as the benchmark to decide on group membership.[15]

The P *Value in Hypothesis Testing*

Inferential statistical tests produce a value for probability. What researchers are interested in is how high or low the chances are (or probability) that the null hypothesis is true. Consider our previous example on severe pain increasing the chances for frequent nighttime awakenings among patients with advanced cancer. If the chances or probability are low, then researchers can decide to reject the null hypothesis. If the chances or probability are high, then researchers will have to stop there; the null hypothesis cannot be rejected.

The probability of the null hypothesis being true has a name; it is called the *P* value, and it ranges from 0 to 1. A *P* value of 0.67 shows 67% probability ($0.67 \times 100 = 67\%$) that the null hypothesis is true, whereas a *P* value of <.002 shows <0.2% probability ($<0.002 \times 100 = <0.2\%$) that the null is true. The exact question that the *P* value answers is this: If the null hypothesis was indeed true, how likely would it be to observe an effect, relationship, or difference as extreme as the one observed here?

The smaller the *P* value, the smaller the chances are that the null hypothesis is true. Equally, the greater the chances are that the alternative hypothesis is true, and as such, the higher the statistical significance of the observed result. If the probability is low or very low, for example, the *P* value is <0.002 (or <0.2% probability), then this means that the null hypothesis does not seem to explain the situation well. As such, the alternative hypothesis seems more plausible. You might ask: Can I now go right ahead and reject the null hypothesis?

No hypothesis test is 100% certain. Because the test is based on probabilities, there is always a chance of making an incorrect decision. When researchers run a hypothesis test, they are likely to make either of the two types of errors discussed previously: type I or type II error (Fig 6). The probability of making a type I error is alpha (ie, to reject a null hypothesis that is true). This is the level of significance researchers set for their hypothesis testing. An alpha of 0.05 indicates that they are willing to accept a 5% chance that they are wrong when they reject the null hypothesis when the null hypothesis is actually true. What researchers do in practice is they compare the *P* value derived from their tests against the alpha. A *P* value that is below alpha (eg, below 0.05) shows low probability that a null hypothesis is true; therefore, it can be rejected. This also implies a statistically significant result, which practically means that a hypothesized effect, relationship, or difference is likely true; in other words, there is evidence it really exists.

Let us go back to our example trial from Pakravan et al[8] in Fig 5. See how the researchers tested for differences between the two groups in terms of age, sex, smoking, and so on. The null hypothesis for all these tests was: "there is no difference between the two groups." The researchers set the significance level alpha to 0.05, meaning that all *P* values below alpha would point toward rejecting the null hypothesis, thus implying that the two groups were indeed different. What you see here is that no *P* value was below the alpha level; the null hypothesis cannot be rejected. The two groups seem to be roughly similar in respect to all of these characteristics. The researchers did the same when they compared pain scores between the two groups at the different time points (see Fig 5). At weeks 0 to 2, the groups were roughly similar in terms of pain scores; no *P* value below 0.05. But see what happens at weeks 3 and 4; *P* values are way below 0.05. If you look at the mean pain scores, you can see a difference of about 2 points on average. The *P* value suggests that there is very low chance for this difference to exist if the null hypothesis was indeed true. As a result, the null hypothesis does not seem to explain the difference very well. The two groups seem to be quite different in terms of pain scores at weeks 3 and 4 (that is quite the opposite of what the null hypothesis suggested). Indeed, the treatment seems to begin to have an effect on patients' perceived pain at weeks 3 and 4.

Confidence Intervals in Hypothesis Testing

Apart from *P* values, confidence intervals can also be used in hypothesis testing. For instance, confidence intervals can be calculated for the difference between the mean scores of two groups. The confidence interval provides information about whether the difference was statistically significant, while it also gives an estimate of the true difference in the wider population. Table 1 provides data from a trial that tested a 12-week resistance training intervention developed for patients on adjuvant radiotherapy for breast cancer.[16] The mean

TABLE 1
Confidence Intervals in Hypothesis Testing.*

	Arm (*N*)	Mean (SD)		Adjusted† mean change (95% CI)	Adjusted† between group difference (95% CI)	*P* value	Effect size *d*
		Before intervention	After intervention				
Total fatigue‡	Exercise (77)	5.9 (2.2)	5.4 (2.3)	-0.5 (-0.9 to -0.2)	-0.5 (-1.0 to -0.0)	.044§	0.25
	Relaxation (78)	6.0 (2.0)	5.9 (1.9)	-0.0 (-0.4 to 0.3)			
Global quality of life‖	Exercise (76)	59 (21)	64 (25)	4.6 (0.1 to 9.2)	3.0 (-3.5 to 9.5)	0.37	0.15
	Relaxation (72)	61 (20)	62 (21)	1.6 (-3.1 to 6.3)			

CI, confidence interval; N, sample size; SD, standard deviation.

* Adapted from Steindorf et al.[14]

† Regression models are adjusted for baseline value.

‡ Fatigue Assessment Questionnaire. Fatigue scores square-root transformed, i.e. they are on a 0-10 scale. Higher scores indicating worse fatigue.

§ Indicates statistical significance at alpha 0.05.

‖ European Organisation for Research and Treatment of Cancer (EORTC QLQ-C30), version 3.0. Scores are on a 0-100 scale. Higher scores indicate better quality of life.

difference in scores before and after the intervention was calculated for two variables in this sample. A 95% confidence interval gives an estimate of true differences in the wider population. From a clinical perspective, we know that, for the variable total fatigue 95 of 100 hypothetical confidence intervals would contain the true difference, and this would lie somewhere between -1.0 and -0.0 (note this value is close to 0 but not 0). Because we are talking about differences in means, a difference in means with an exact value of 0 should be interpreted as no difference whatsoever between the two groups. Note how for the variable global quality of life, the 95% confidence interval goes from -3.5 to 9.5. This means that a 0 value is included here; you can expect that this result is not statistically significant. The *P* value of 0.37 that is attached to this test shows that at alpha 0.05, the result is not significant. The clinician understands that there is no clarity regarding a difference between the two groups in relation to this variable. If 100 groups were sampled, then in some cases the exercise group would do better; in other cases, the relaxation group would do better, and yet in other cases, there would be no difference at all. Therefore, the evidence is not really convincing about any effects on global quality of life.

Data Interpretation

With data analysis completed, the most interesting and rewarding part (and at the same time most difficult) is the art of interpretation of the emerging evidence. Evidence without interpretation is the exact same thing as raw data without analysis; they are of no use to anyone. When you read papers, you want the authors to have written an intriguing and thought-provoking discussion of their results that motivates the reader to think more broadly about the importance of the evidence and its potential uses.

One key thing to remember is that, although statistics test hypotheses to indicate that an effect, relationship, or difference is real (and not due only to chance), statistically significant results may not always be clinically important. Statistical significance usually is a function of the sample size. The larger the sample, the easier it is to show a statistically significant result even though the result might not have any real meaning for use in practice.

Effect Sizes in Data Interpretation

Jacob Cohen, one of the most influential statisticians of the 20th century, wrote in 1990 that "the primary product of a research inquiry is one or more measures of effect size, not *p*-values."[17] Indeed, *P* values are a good introduction to the world of inferential statistics. However, results of inferential statistical tests should be described in terms of *measures of magnitude*, i.e. not just whether or not a treatment benefits patients but how much it benefits them (if at all).[18] *P* values will inform us about whether an effect, relationship, or difference might exist in reality. Measures of magnitude will tell us how small or large this effect, relationship, or difference is; this information can be clinically useful for decision-making.

Measures of magnitude come in the form of effect sizes.[19] Effect sizes can be calculated for any type of association or comparison and, therefore, may refer to differences in mean scores, differences in odds, or the size of correlation between variables. Effect sizes can be absolute when the variable under investigation has intrinsic meaning. Consider this example. Cognitive-behavioral therapy for insomnia (CBT-I) is tested as part of a pilot trial to see whether it is related to gains in total sleep time at night (number of hours) in patients with cancer.[20] The results show a statistically significant difference on total sleep time between the intervention group (CBT-I) and control group (usual care); let us say the *P* value is 0.02. The variable total sleep time has intrinsic meaning (ie, number of hours). An absolute effect size can be calculated for the difference in total sleep time between intervention and control group; let us say the effect size is 1.6 hours. On average, the intervention group gained 1.6 hours of extra sleep time after CBT-I compared to the control group (eg, mean total sleep time of 6.8 hours for the intervention group − mean total sleep time of 5.2 hours for the control group = 1.6 hours). This absolute effect size can communicate important information to clinicians and patients, and it has concrete meaning: an average of 1.6 hours can easily be considered for its clinical importance.

Imagine also that patients in both groups were asked to self-assess their sleep quality on a 0-10 visual analogue scale. An absolute effect size could be calculated to show an average improvement of 2.3 points on the scale in favor of the intervention group. Although this is good information, we are unsure whether a 2.3 change is large enough or trivial. The variable has no intrinsic meaning, or it can be difficult to express how much change is clinically important. A standardized effect size can be calculated to take into account variability in the measured improvement. In other words, a standardized effect size also looks at the standard deviation (variability) of the data in the variable and not just the absolute size of the difference. The larger the variability, the smaller the standardized effect size because the direction of the effect becomes diluted. Standardized effect sizes are unitless; this makes it easier to compare effect sizes that come from different studies or where different measures were used to measure the same variable.

Cohen's *d* probably is the standardized effect size we are most familiar with.[21] Standardized effect sizes are interpreted against set benchmarks or rules of thumb. For instance, Cohen[21] classified effect sizes as small (d = 0.2-0.49), medium (d = 0.5-0.79), and large ($d \geq 0.8$). Suppose that in our example above the standardized effect size is calculated as 0.39; this points to the direction of a small (although not too small) effect size of CBT-I on sleep quality. Similarly, in Table 1,[16] a *P* value of 0.044 indicates a statistically significant difference between exercise (intervention group) and relaxation (control group); that is, a true difference might exist in reality between the two groups. However, an effect size of d = 0.25 implies only a small size of the effect of exercise training on patients' total fatigue; in other words, from a clinical perspective, the effect might not be as important.

Critical Thinking in Data Interpretation

Let us consider another example (Table 2). Suppose you have three independent, fictitious trials that test a new medication against what currently is standard practice for the treatment of prostate cancer. The first trial has a sample size of 10,000 patients and concludes with a statistically significant effect of the new medication, whereby the new medication reduces prostate-specific antigen (PSA) levels by 0.5% on average; this is the absolute effect size. Statistical significance is confirmed with an extremely low *P* value. However, a reduction of 0.5% is negligent to justify approval of the new medication for use in clinical practice, particularly when side effects and costs are also taken into account. Because of the very large sample size, the researchers were still able to show statistical significance of a rather minimal effect.

The second trial involves only 10 patients. Results show that the new medication is associated to a 30% reduction in PSA levels, a quite large effect size. However, the result is not statistically significant despite being obviously clinically important. The small sample size does not allow researchers to show statistical evidence that this is not a chance finding. As such, the new medication possibly will not be used in clinical practice unless more and larger trials are done to replicate this same finding. Finally, the third trial shows an average 20% reduction in PSA levels, which is both clinically important and statistically significant. Can this result lead to the new medication being approved for use in clinical practice? Probably yes, particularly if similar studies produce similar results.

TABLE 2
Hypothetical Scenarios Showcasing Statistical Significance Versus Clinical Importance.

	New medication tested against standard treatment	Tested in	*P* value	Statistical significance (ie, <0.05)?	Clinical importance?	Possible use in practice?
Trial 1	Reduces PSA levels by 0.5% on average	10,000 patients	.00003	Yes	No (small effect size)	Uncertain but probably not. Do the side effects and cost justify the use?
Trial 2	Reduces PSA levels by 30% on average	10 patients	.74	No	Yes (large effect size)	Uncertain but possibly not. Can it help many people? More research is required.
Trial 3	Reduces PSA levels by 20% on average	200 patients	.004	Yes	Yes (moderate effect size)	Probably yes, particularly if these effects are replicated in similar studies.

PSA, prostate-specific antigen.

Conclusion

Data processing in quantitative research involves the combination of careful data management techniques, knowledge of statistics, and critical thinking skills to aid interpretation. Developing capacity in the management, analysis, and interpretation of quantitative research data can have a multifaceted impact in enhancing nurses' confidence in understanding, evaluating, and applying quantitative evidence in cancer nursing practice. Throughout this article, we have described using practical examples, the stages of quantitative data processing. We have also looked to clarify some concepts to support both students and researchers.

Having considered several important aspects in this article, several functions can be considered as advantages in the processing of quantitative research data. Statistics can deal with large numbers of data, variables, and samples. They can quantify the effect of a new treatment, service, or intervention. They can also explore relationships between two variables while controlling for third ones. Statistics can help summarize characteristics of the sample for possible generalization to the wider population. Using statistics, the analysis can be replicated using the same data set. Personal bias is avoided via careful data management and critical thinking to help researchers keep a distance from the data, while casting a critical eye on them.

Processing quantitative research data is not without its challenges. Data management can be time-consuming and requires skillful analysts. The quality of the results depends on the quality of the data. Several points could be made about the effects of missing data or imputation techniques, about data derived from less well validated measures, or indeed appropriate for the research. While statistics produce results to prompt consideration for clinical practice or future research, some results can be difficult to interpret or explain and, as such, difficult to apply to the real world. Errors in statistical analyses may return incorrect results and misleading conclusions. Statistical significance does not always translate into clinical importance, and overreliance to statistical significance might overlook potentially important hints toward important discoveries.

References

1. Watson R. Quantitative research. *Nurs Stand*. 2015;29(31):44–48. https://doi.org/10.7748/ns.29.31.44.e8681.
2. Polit DF, Beck CT. *Essentials of Nursing Research: Appraising Evidence for Nursing Practice*. 9th ed Philadelphia, PA: Lippincott Williams and Wilkins; 2018.
3. van den Berg RG. Measurement Levels – What and Why? SPSS Tutorials. https://www.spss-tutorials.com/measurement-levels/. Accessed February 22, 2023.
4. Van Den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*. 2005;2(10):e267. https://doi.org/10.1371/journal.pmed.0020267.
5. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402–406. https://doi.org/10.4097/kjae.2013.64.5.402.
6. Anderson G, Ebadi M, Vo K, Novak J, Govindarajan A, Amini A. An updated review on head and neck cancer treatment with radiation therapy. *Cancers (Basel)*. 2021;13(19):4912. https://doi.org/10.3390/cancers13194912.
7. Kotronoulas G, Papadopoulou C, Simpson MF, McPhelim J, Mack L, Maguire R. Using patient-reported outcome measures to deliver enhanced supportive care to people with lung cancer: feasibility and acceptability of a nurse-led consultation model. *Support Care Cancer*. 2018;26(11):3729–3737. https://doi.org/10.1007/s00520-018-4234-x.
8. Pakravan F, Ghalayani P, Emami H, Isfahani MN, Noorshargh P. A novel formulation for radiotherapy-induced oral mucositis: Triamcinolone acetonide mucoadhesive film. *J Res Med Sci*. 2019;24:63. https://doi.org/10.4103/jrms.JRMS_456_18.
9. Messori A, Caccese E, D'Avella MC. Estimating the 95% confidence interval for survival gain between an experimental anti-cancer treatment and a control. *Ther Adv Med Oncol*. 2017;9(11):721–723. https://doi.org/10.1177/1758834017731084.
10. Darling HS. Are you confident about your confidence in confidence intervals? *Cancer Res Stat Treat*. 2022;5(1):139–144.
11. Thompson B. Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychol Sch*. 2007;44(5):423–432. https://doi.org/10.1002/pits.20234.
12. Yarandi HN. Hypothesis testing. *Clin Nurse Spec*. 1996;10(4):186–188.
13. Shreffler J, Huecker MR. Types of variables and commonly used statistical designs. *StatPearls [Internet]*. Treasure Island, FL: StatPearls Publishing; 2022.
14. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth*. 2019;22(1):67. https://doi.org/10.4103/aca.ACA_157_18.
15. Drury A, Payne S, Brady AM. Identifying associations between quality of life outcomes and healthcare-related variables among colorectal cancer survivors: a cross-sectional survey study. *Int J Nurs Stud*. 2020;101: 103434. https://doi.org/10.1016/j.ijnurstu.2019.103434.
16. Steindorf K, Schmidt ME, Klassen O, et al. Randomized, controlled trial of resistance training in breast cancer patients receiving adjuvant radiotherapy: results on cancer-related fatigue and quality of life. *Ann Oncol*. 2014;25(11):2237–2243. https://doi.org/10.1093/annonc/mdu374.
17. Cohen J. Things I have learned (so far). *Methodological Issues & Strategies in Clinical Research*. Washington, DC: American Psychological Association; 1992:315–333. https://doi.org/10.1037/10109-028.
18. Kline RB. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: American Psychological Association; 2006. https://doi.org/10.1037/10693-000.
19. Peterson SJ, Foley S. Clinician's guide to understanding effect size, alpha level, power, and sample size. *Nutr Clin Pract*. 2021;36(3):598–605. https://doi.org/10.1002/ncp.10674.
20. Garland SN, Johnson JA, Savard J, et al. Sleeping well with cancer: a systematic review of cognitive behavioral therapy for insomnia in cancer patients. *Neuropsychiatr Dis Treat*. 2014;10:1113–1123. https://doi.org/10.2147/NDT.S47790.
21. Cohen J. Statistical power analysis for the behavioral sciences. *Stat Power Anal Behav Sci*. 1988;2nd:567. https://doi.org/10.1234/12345678.