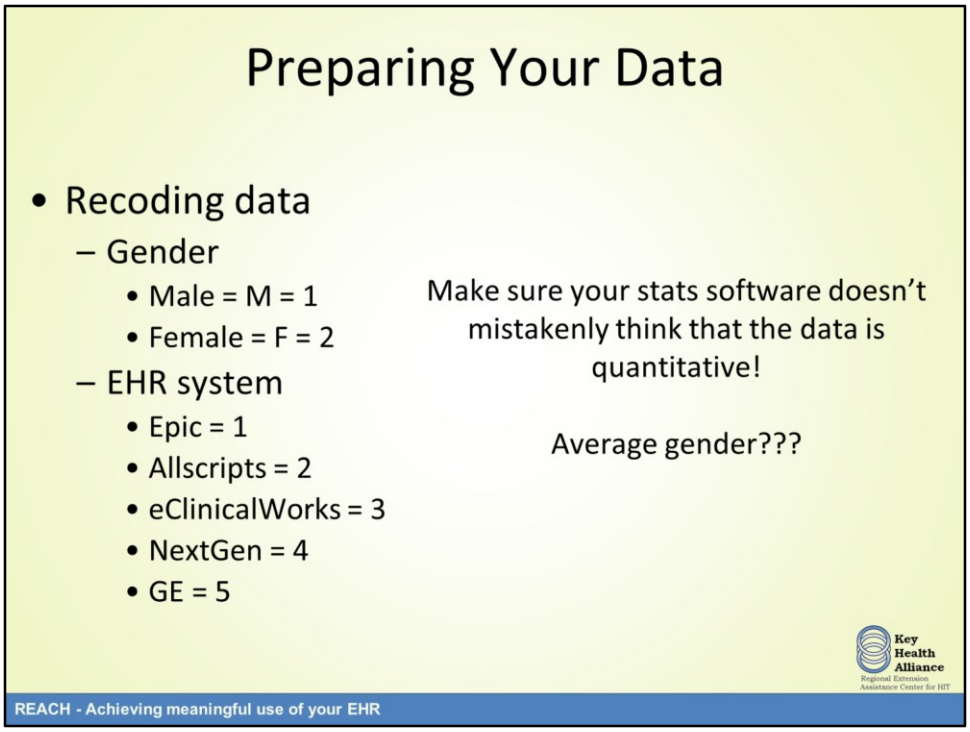
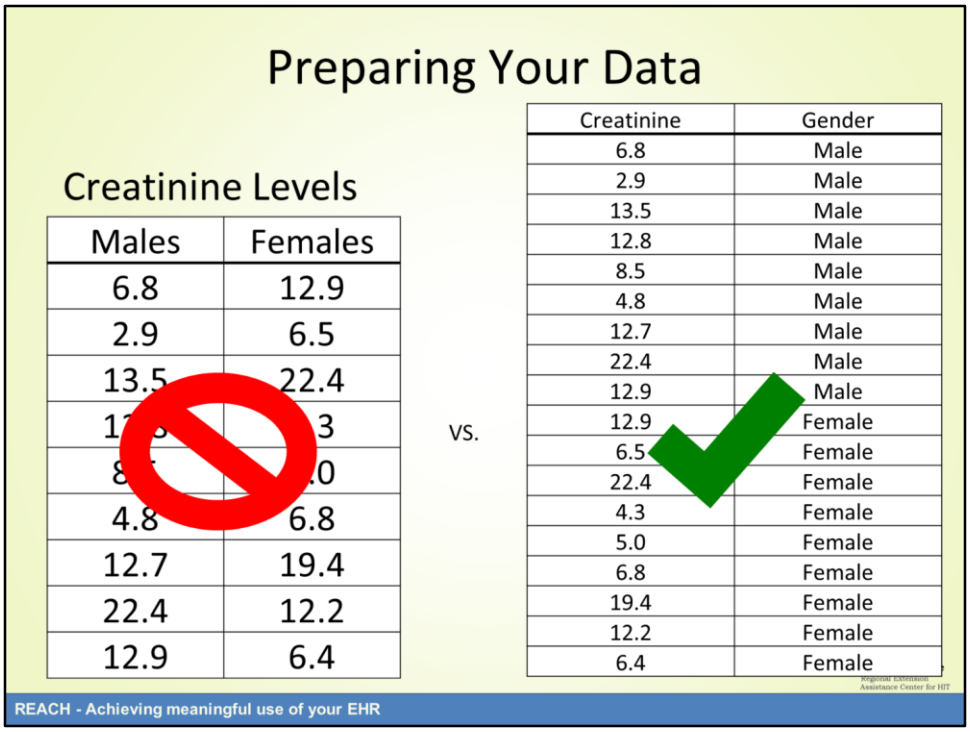
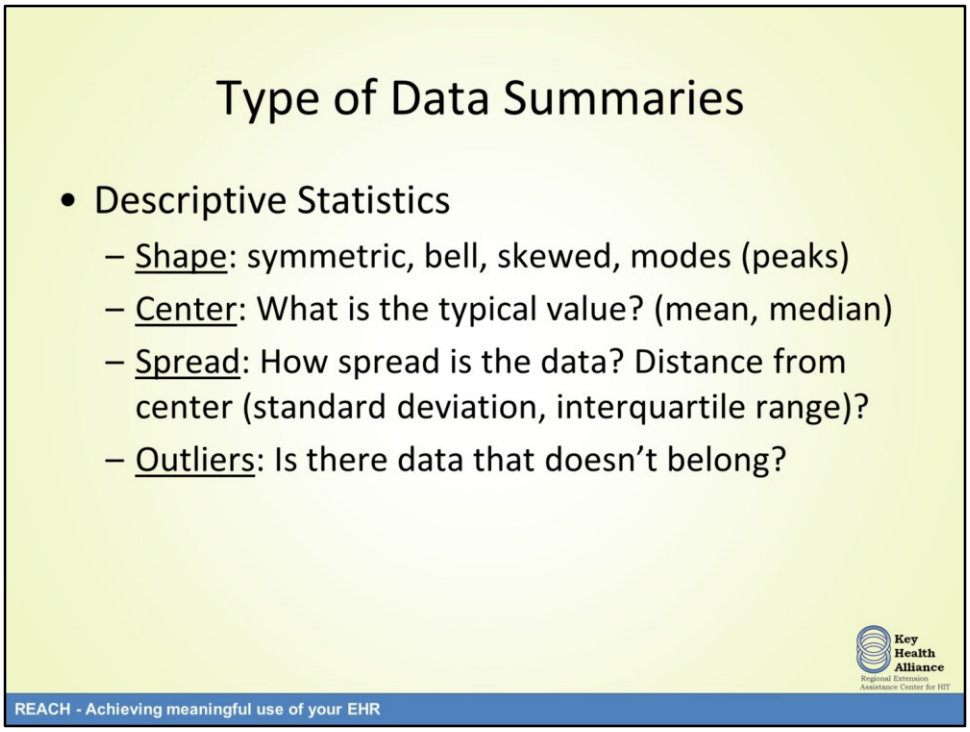
Welcome to the Data Analytics Toolkit PowerPoint presentation on basic analytics. In this presentation, you will be introduced to the application of basic analytic methods to healthcare data and the implications for meeting the demands of meaningful use.



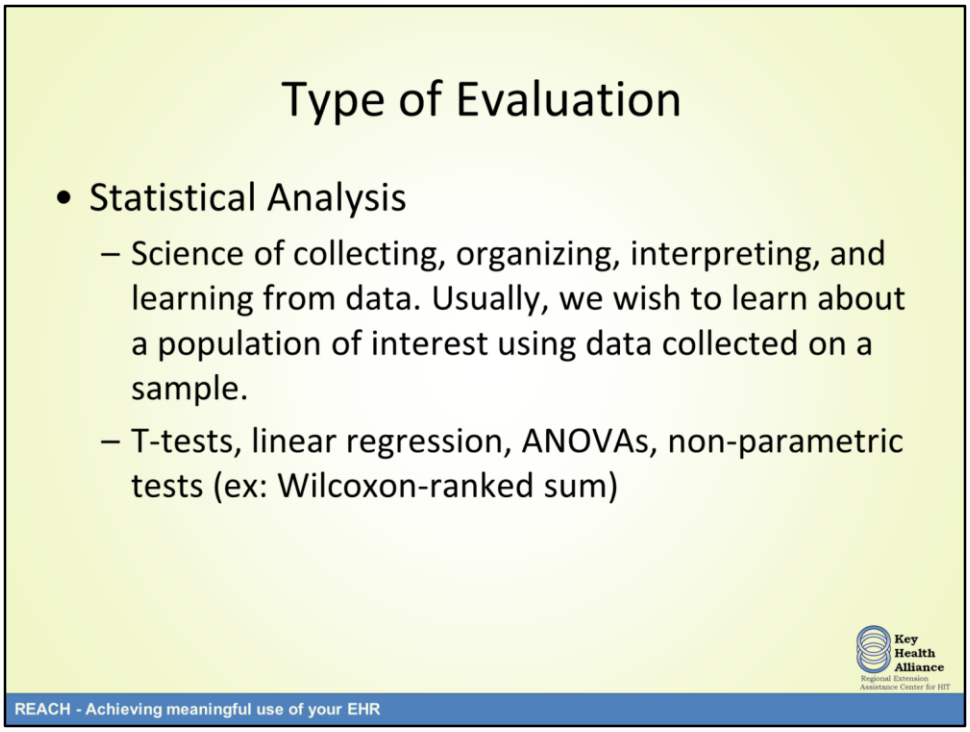
The first thing you must consider when you take on an analytics project is how your data is prepared. Often times you will encounter data that is coded. For instance, Gender may be coded numerically where 1 is equal to males and 2 is equal to females. You might have other types of data that are coded, such as that from EHR systems. The data dictionary will provide you the insights into deciphering what each of the numbers in your dataset refer to. However, when you are using coded data for analysis purposes, you often have to tell your stats software that the data is not to be interpreted as a number. Otherwise, what the software may do is provide you with descriptive stats that don't make sense. For instance, does it make sense to talk about average gender? Not really.



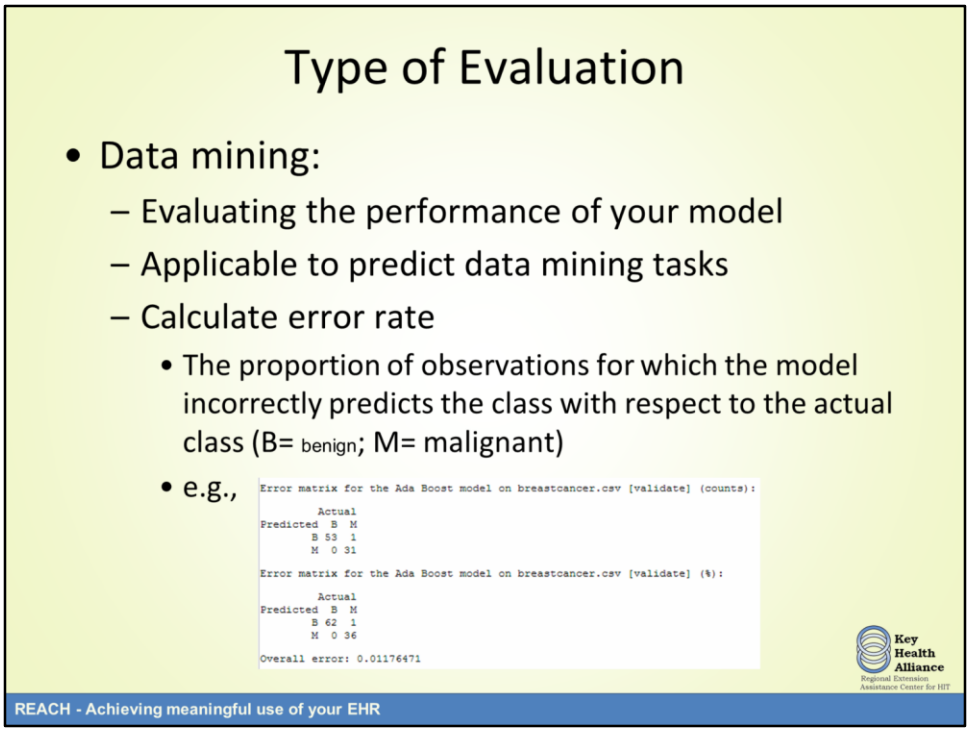
You must also consider the format of your data. Sometimes you don't have control over the format because you are at the mercy of the data storage methods that were adopted. However, you may encounter datasets that are organized incorrectly, leaving you scratching your head on what the data actually mean. For instance, let's say you had a dataset that included Creatinine levels for males and females. The data could be arranged where there is a column for males and a column for females. However, the issue with this arrangement is that there isn't any indication as to what was measured. That is, how do you know that those numbers represent creatinine levels? An alternative arrangement is to have two columns of data that represent each of the attributes: creatinine and gender. Each row of data represents a single person and the creatinine level and gender for that person are shown. This arrangement is much clearer and also easier for analytic software to interpret.



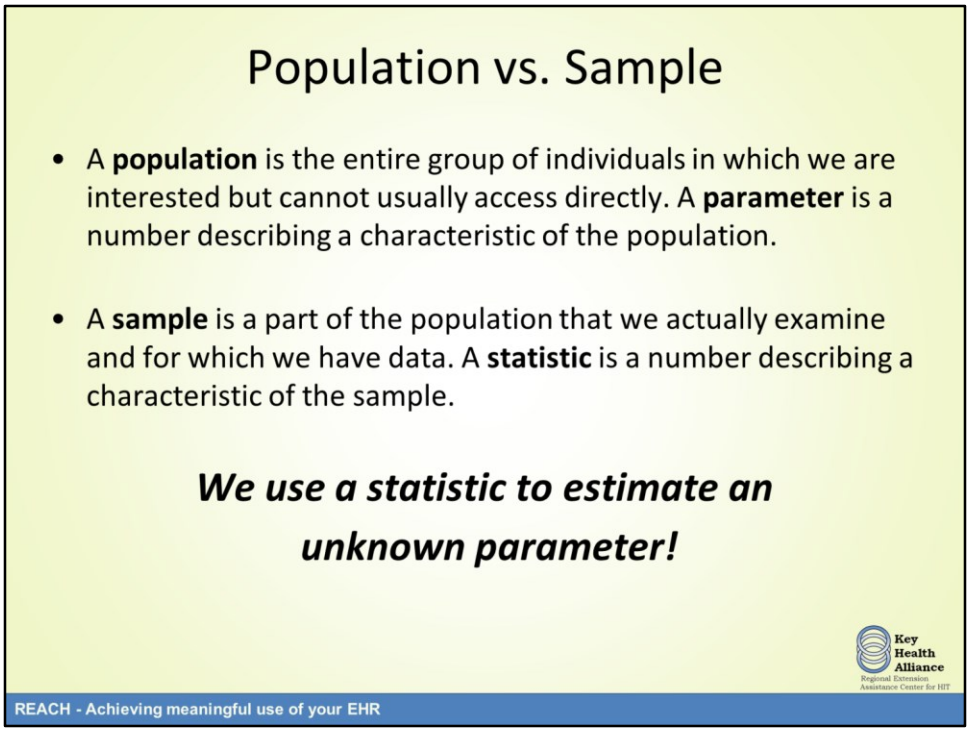
Quantitative data should always be described in terms of shape, center, spread, and outliers. Shape refers to the symmetry of the data. Typically, we want to achieve a bell shaped distribution where data is normally distributed. Data can also be skewed or have multiple modes (which are peaks in the distribution). A histogram is typically used for evaluating the shape of the data. Center can be described by what the typical value is. Mean and median are ways for showing center. Spread describes how variable the data is, or how distant the extremes are from the center. Standard deviation, variance, and interquartile ranges are ways of describing spread. Outliers are data points that are too extreme to be considered reliable, that is, data that doesn't belong in your dataset. Outliers can be identified with various methods, such as boxplots, IQRx1.5, or other methods, however, it is really up to the researchers to determine if a value is an outlier.



There are various evaluation methods that can be adopted for analyzing data. Statistics is one such method. Statistics is the science of collecting, organizing, interpreting, and learning from data. Usually we wish to learn about a population of interest using data collected on a sample. There are many different types of statistical procedures that can be adopted depending on the type of data you are looking at and the questions that you pose. Some of the most popular statistical procedures include T-tests, linear regression, ANOVAs, and non-parametric tests such as the Wilcoxon--ranked sum test. Statistical evaluations are driven by hypotheses - you state a hypothesis and test the truth of that hypothesis.

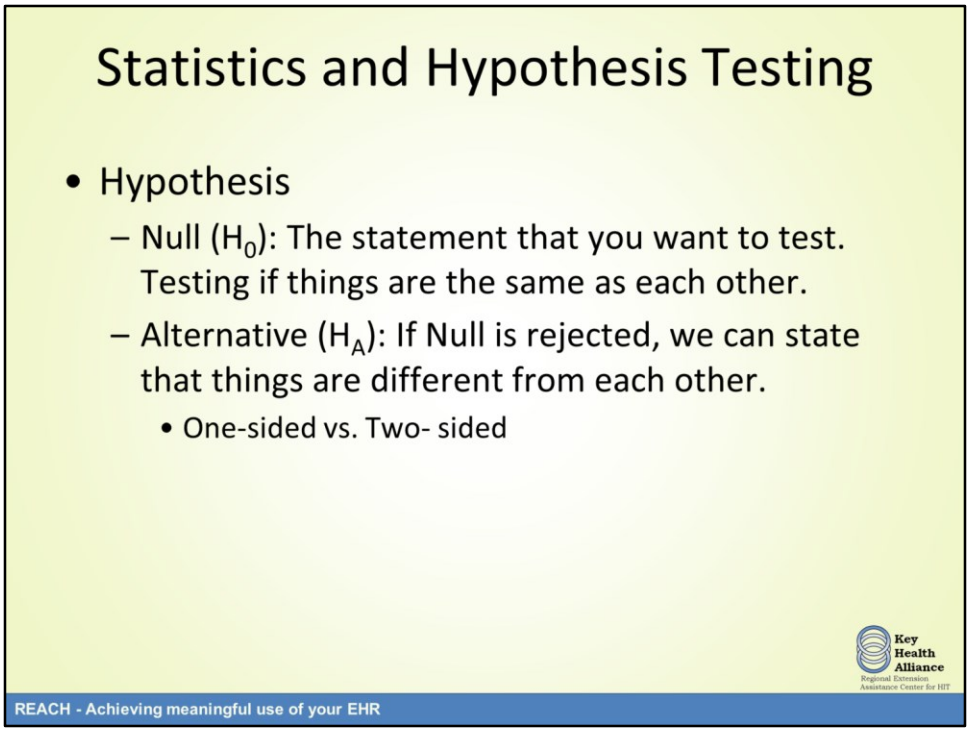


Another evaluation method is data mining, which is not based off of hypothesis testing. The purpose of data mining is to discover new information from data. Data mining methods can be used for describing data or predicting outcomes. The data mining methods construct a model and that model can be applied to new data. The performance of the model is determined using an error rate. The error rate is the proportion of observations for which the model incorrectly predicts the class with respect to the actual class. For instance, if you were attempting to predict if a tumor was benign or malignant using historical data, you can determine the error rate of your model for when it determines that a tumor is benign when in fact it is malignant and when it predicts a tumor is malignant when in fact it is benign. The goal is to construct a model with a low error rate.



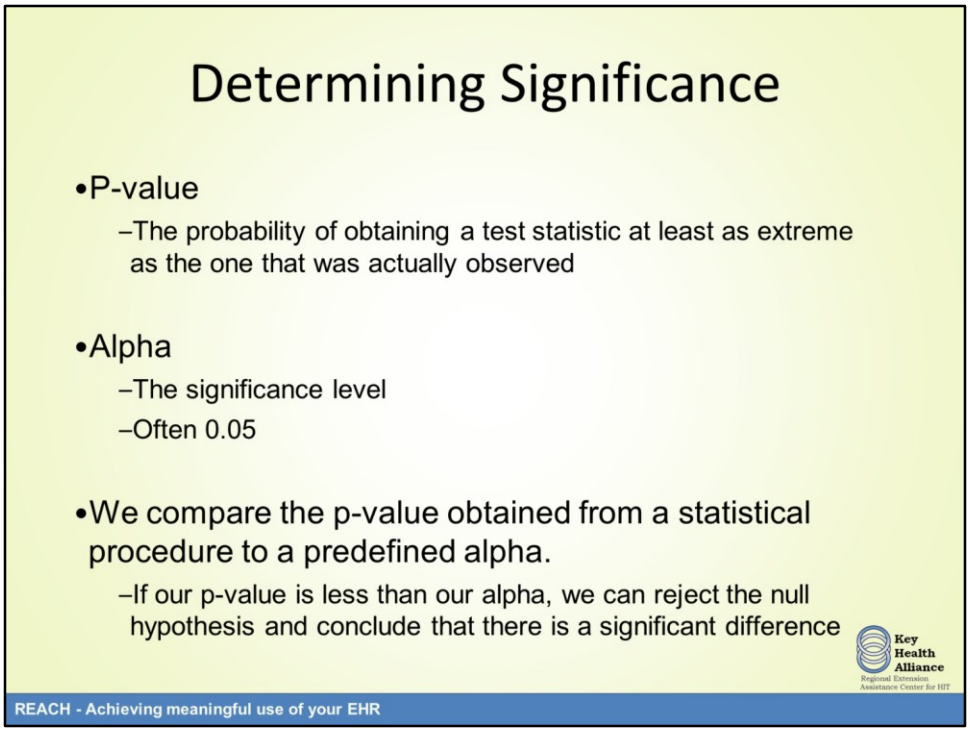
When conducting an analysis, you are typically working with a sample and not a population. This is particularly important for statistical evaluations. A population is the entire group of individuals in which we are interested but cannot usually access directly. A parameter is a number describing a characteristic of the population. An example of getting information about a population is a census. This is one of the only ways of getting information about an entire population. A sample, on the other hand, is a part of the population that we actually examine and for which we have data. A statistic is a number describing a characteristic of the sample.

The purpose of a statistic is to estimate an unknown parameter.



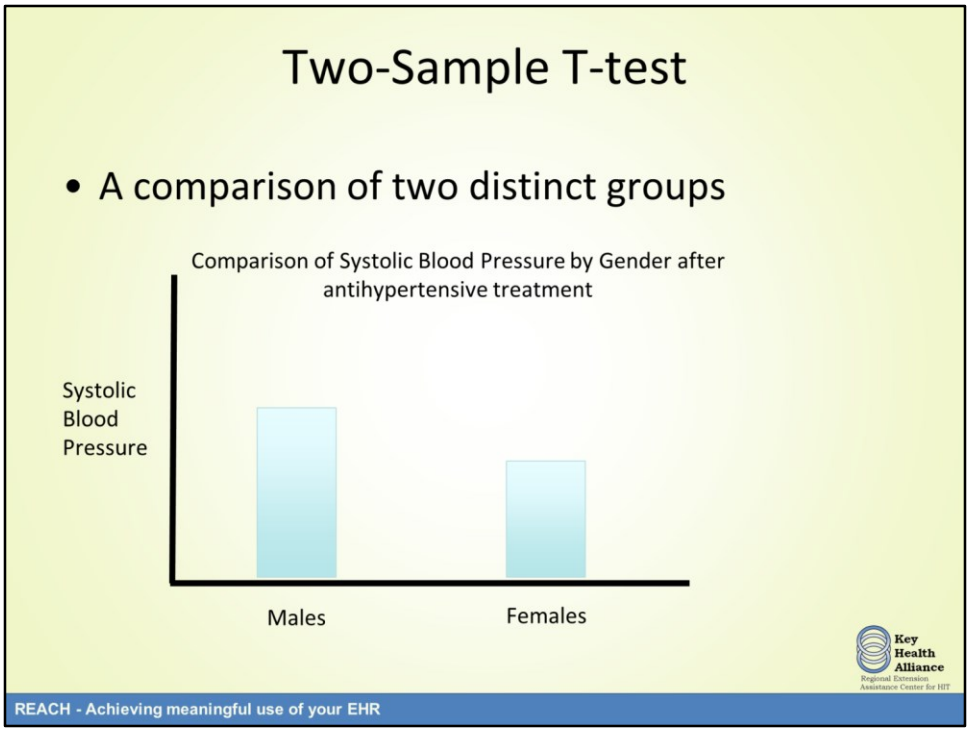
Statistical evaluations require that you state your hypothesis prior to analyzing data. When stating your hypothesis, there exists the null and alternative. The null hypothesis is actually what the statistical procedure is testing. With statistics we are testing if things are the same as each other. The alternative hypothesis is accepted if the null is rejected and we can state that things are different from each other. For many procedures, there are two types of alternative hypotheses. These are important to differentiate because they can significantly change the results of the statistical procedure. A one-sided hypothesis provides directionality for a relationship. For instance, a one sided alternative hypothesis would be, “Patients taking Drug A have lower blood pressure than patients taking placebo” or “Females have higher test scores than males”. A two-sided alternative hypothesis does not provide directionality. For instance, “Patients taking Drug A have different blood pressure levels than patients taking placebo” or “Females have different test scores than males”. The advantage of the two-sided hypothesis is that we do not need to make a prediction regarding the directionality. If we incorrectly predict directionality such as “Females have higher scores than males” but in reality “Males have higher scores than females” we may actually see no significance. However, if we were to have used a two-sided hypothesis, we would probably detect a

difference. By default, we typically stick to a two-sided hypothesis. Yet, in certain circumstances you may use a one-sided hypothesis.

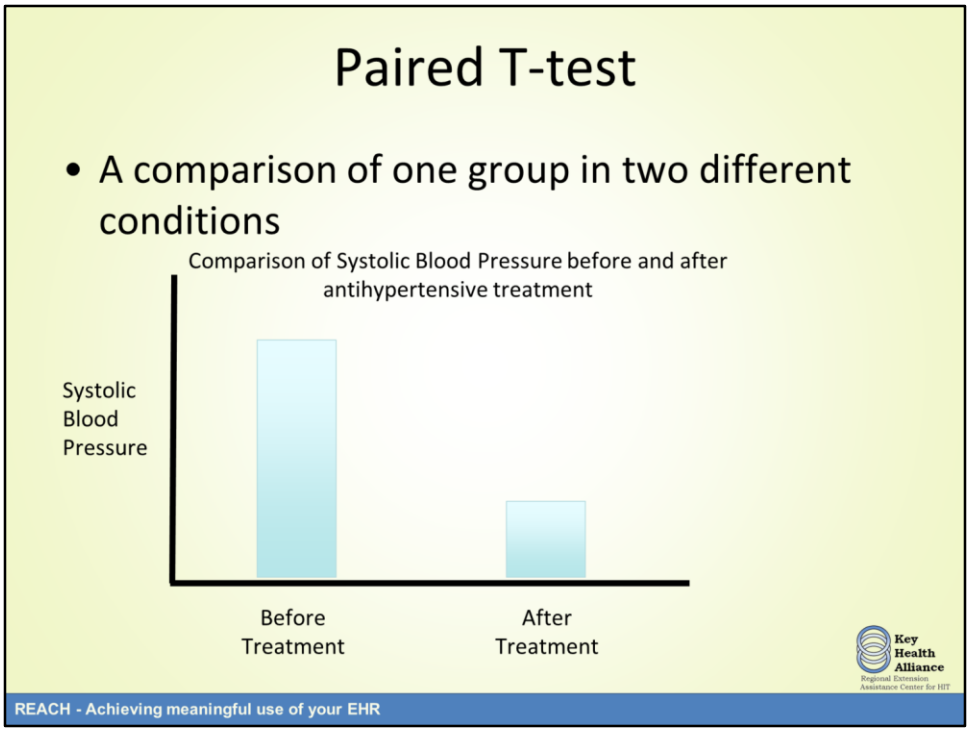


A statistical procedure produces a test statistic which can be converted to a P-value for determining significance. By definition a p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. A researcher will often "reject the null hypothesis" when the p-value turns out to be less than a certain significance level, known as alpha. Alpha is often set at 0.05.

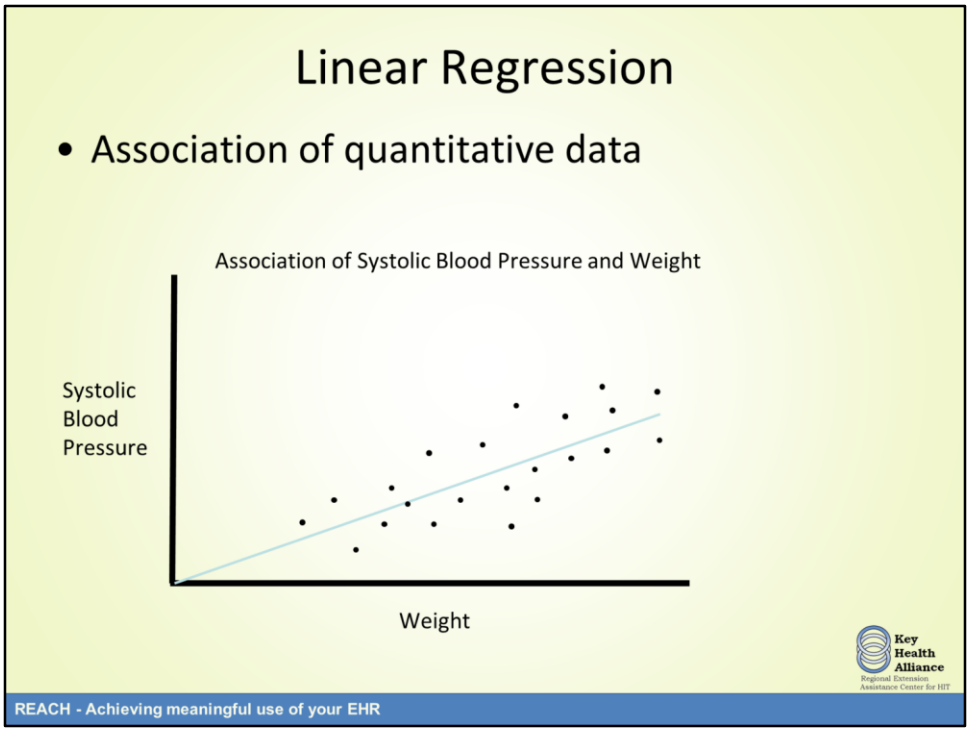
Let's say I wanted to compare average cost of a specific procedure between two hospitals. I could use a statistical procedure to determine if the average cost is significantly different. If after running my statistical procedure, I obtained a p-value of 0.03 and determined this to be less than my alpha of 0.05, I can conclude that there is a significant difference in average costs between the two hospitals. Because the p-value is used to determine if an observed statistic is significant, it is the most important result for interpreting statistical procedures.



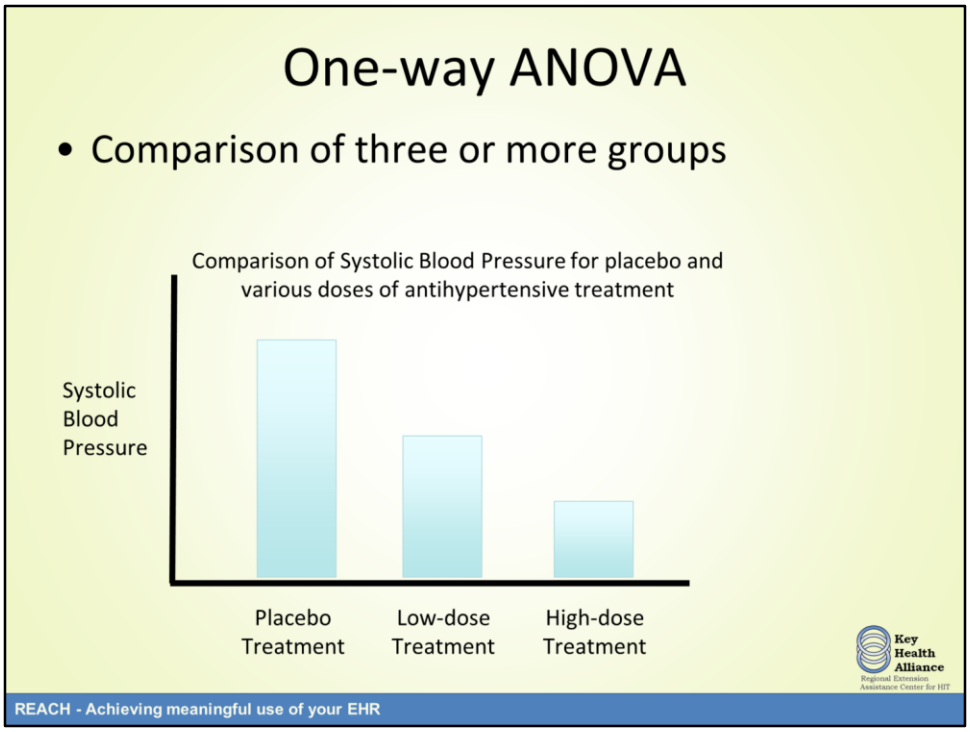
One common statistical method is a two-sample T-test which compares average responses in two distinct groups. The responses in each group are then independent of those in the other group. If we wanted to compare average systolic blood pressure in males and females, our null hypothesis for a two-tailed two-sample t-test would be that the male subjects have equal systolic blood pressure to female subjects. Our alternative hypothesis is that male subjects do not have equal systolic blood pressure to female subjects. As you can see, the male group has higher systolic blood pressure than the female group. We could use a two-sample T-test to compare mean systolic blood pressure plus or minus the standard error of the mean between these two groups. If we were to compare the means plus or minus the standard error of the means, we might see that there is very little crossover between our error bars. Therefore, it is likely that the two-sample t-test would show a p-value less than our alpha of 0.05 thereby establishing that we can reject the null hypothesis and conclude that the male group has significantly higher systolic blood pressure levels than the female group.



If we wanted to compare the same group in two different conditions, you can adopt a paired t-test. Let's say we wanted to compare the mean difference in systolic blood pressure before and after antihypertensive treatment. We could use a paired T-test and evaluate the mean differences in systolic blood pressure before and after treatment. Our two-tailed null hypothesis would be that the mean difference in systolic blood pressure before and after treatment is equal to 0 while our alternative hypothesis would be that the mean difference in systolic blood pressure before and after treatment is not equal to 0. Looking at the means of each group, it is likely that we would calculate a difference for patients, that systolic blood pressure decreases after treatment. If we find a p-value that is less than our alpha we could conclude that after using antihypertensive treatment there is a significant drop in systolic blood pressure.

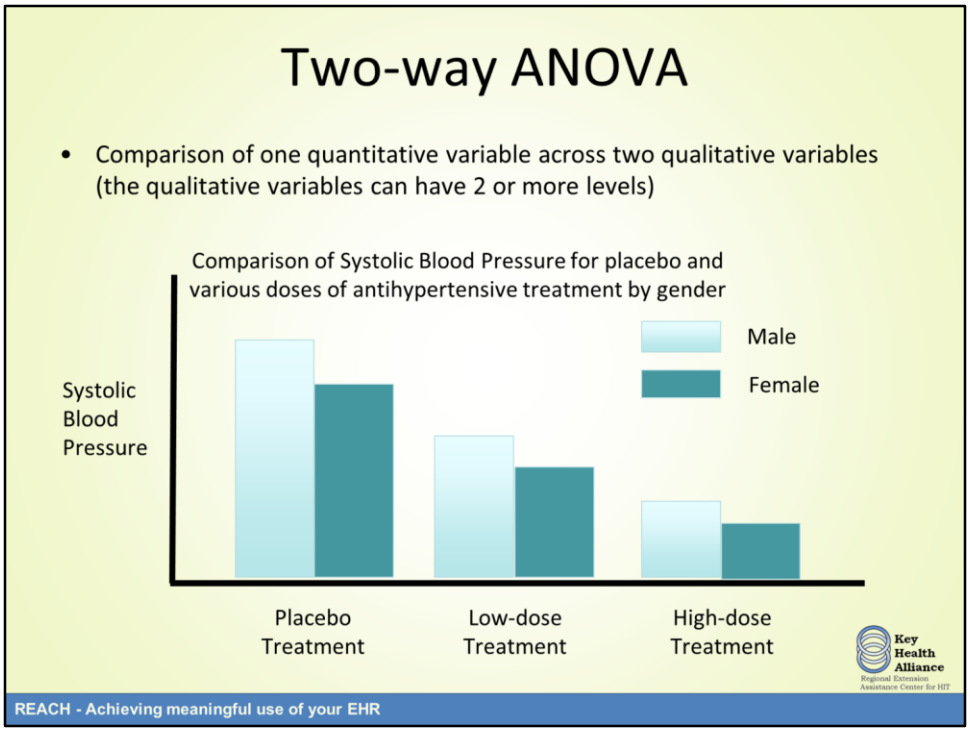


If you wanted to evaluate the association between two quantitative variables, a simple linear regression would be appropriate to adopt. In this example, the association between systolic blood pressure and weight were tested. The null hypothesis for a simple linear regression is that the slope of the best fit line is equal to zero, this would occur if the line was flat. The alternative hypothesis is that the slope for the best fit line is not equal to zero. A correlation coefficient, represented by the r-squared value would then show the strength of the relationship. The closer the r-squared value is to 1 or -1, the stronger the relationship. In this case, let's assume our p-value was less than our alpha of 0.05, therefore we can reject the line hypothesis and conclude that there is a relationship between systolic blood pressure and weight. If our r-squared value is 0.5, this would mean that there is a moderate strength for the observed association.

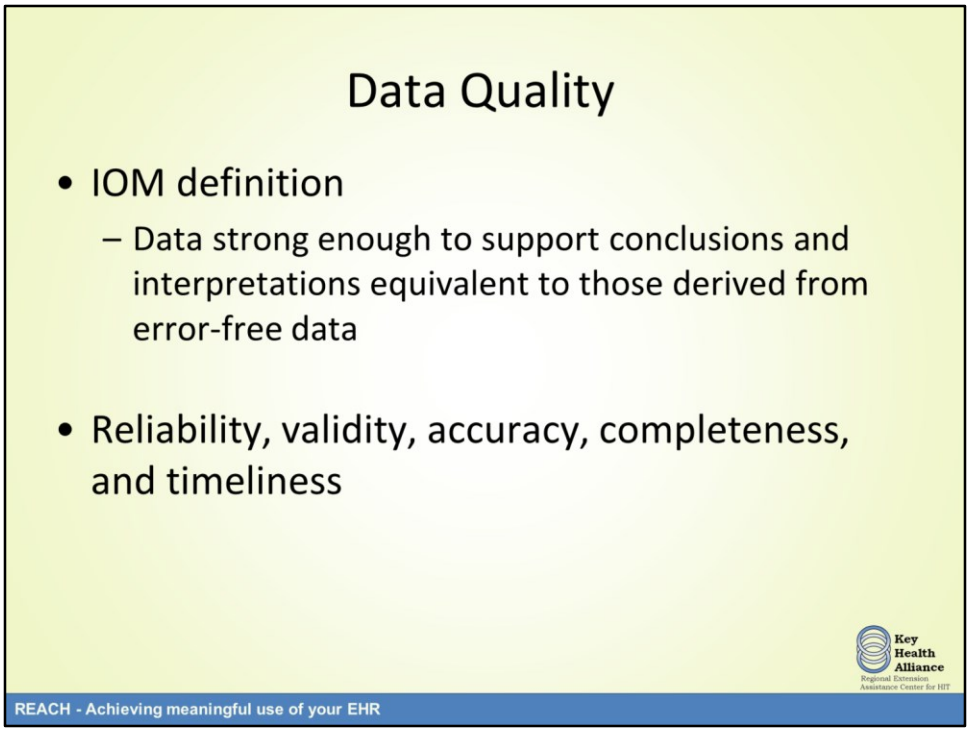


When you want to compare the mean of three or more groups, a one-way ANOVA is the best test to adopt. Let's say you wanted to compare mean systolic blood pressure in subjects that either took a placebo, low-dose, or high-dose of an antihypertensive treatment. We can use a one-way ANOVA to determine if there are significant differences in the average systolic blood pressure between the three drug conditions. The null hypothesis would be that the mean systolic blood pressure is the same for each treatment while the alternative hypothesis is that at least one treatment group is different than at least one other treatment group in terms of systolic blood pressure. We would obtain an F score to determine the p-value and if our p-value is less that our alpha we can reject the null hypothesis and conclude that at least one treatment group has significantly different systolic blood pressure than at least one other treatment group. However, we would have to use a Tukey post hoc test to look at the pair-wise comparisons and determine which group had a significant difference than at least one other group. It is likely that the Tukey would find the placebo and high-dose groups have significantly different systolic blood pressure; therefore, we can conclude that the high-dose treatment leads to significantly lower systolic blood pressure levels than the

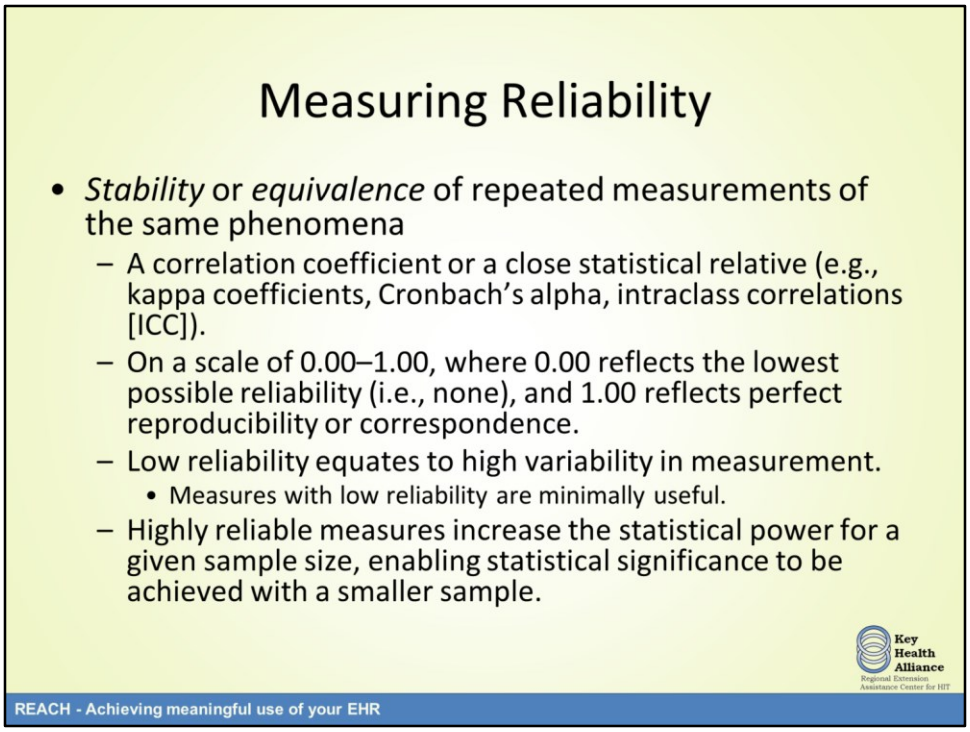
placebo group.



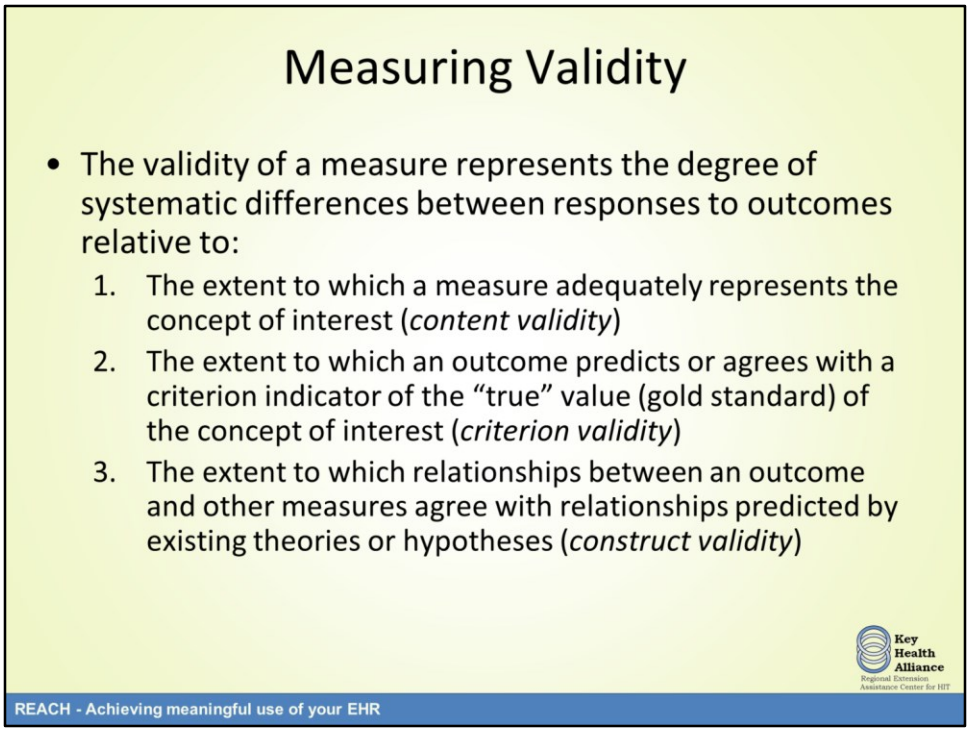
In this example we want to compare mean systolic blood pressure in men and women after different types of treatments. There were three treatments, a placebo, low-dose, and high-dose. Thus, our independent variables are treatment and gender while our dependent variable is systolic blood pressure. We can use a two-way ANOVA to determine if the treatment and gender had a main effect or if there is an interaction between treatment and gender on systolic blood pressure. Our null hypotheses would be that the mean systolic blood pressure are equal for all the treatments, mean systolic blood pressure are equal for both genders, and that there is not an interaction between the treatment and gender on the mean systolic blood pressure. If we look at the results of our study, we would most likely discover that the mean systolic blood pressure does differ between each of the treatments and genders where females have lower levels than males. It is unlikely, however, that an interaction exists because there isn't a significant deviation where one type of treatment for one gender group has a significantly different effect.



The quality of the data ultimately influences the value of the conclusions that are made from any type of evaluation. The Institute of Medicine defines data quality as “Data strong enough to support conclusions and interpretations equivalent to those derived from error-free data”. Data quality can be measured based on the reliability, validity, accuracy, completeness, and timeliness of the data.



Reliability is synonymous with stability, or equivalence of repeated measurements of the same phenomena. One way to measure reliability is by calculating a correlation coefficient of similar statistic such as a kappa coefficient, cronbach's alpha, or intraclass correlations. They are typically on a scale of 0 to 1, where 0 reflects the lowest possible reliability (that is none) and 1 reflects perfect reliability or correspondence. Low reliability equates to a high variability in measurement. Thus, measures with low reliability are minimally useful. Highly reliable measures increase the statistical power for a given sample size, enabling statistical significance to be achieved with a smaller sample.



Validity is a measure that represents the degree of systematic differences between response to outcomes relative to several factors. One factor, known as content validity is the extent to which a measure adequately represents the concept of interest. For example, how well do blood thyroid hormone levels represent thyroid function? Next, criterion validity refers to the extent to which an outcome predicts or agrees with a criterion indicator of the true value of the concept of interest. Does the measure's thyroid hormone level for a patient actually agree with the true level in that patient? Finally, construct validity refers to the extent to which relationships between an outcome and other measures agree with relationships predicted by existing theories or hypotheses. For example, we know the normal range or thyroid hormone levels allowing us to compare the measured thyroid hormone level for a patient relative to the accepted normal range.

Data Quality

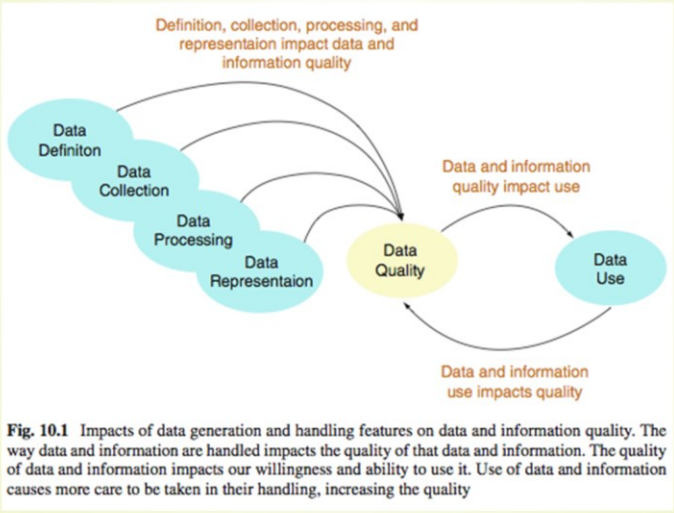
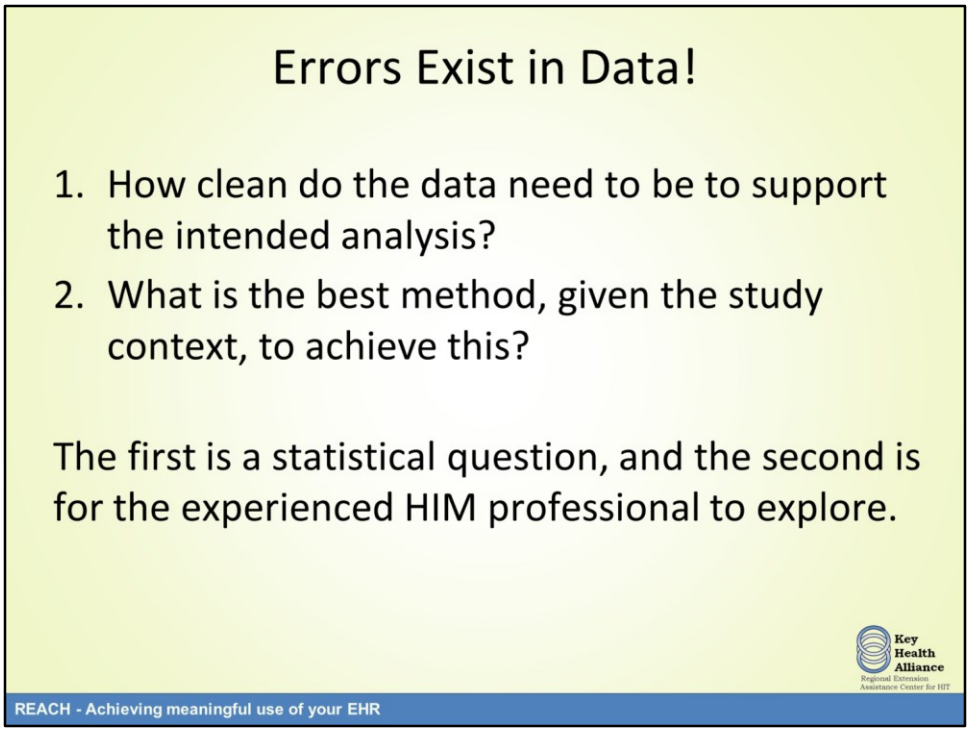


Fig. 10.1 Impacts of data generation and handling features on data and information quality. The way data and information are handled impacts the quality of that data and information. The quality of data and information impacts our willingness and ability to use it. Use of data and information causes more care to be taken in their handling, increasing the quality

Key Health Alliance
Regional Extension Assistance Center for HIT

Various factors can influence data quality. The way in which the data is defined, collected, processed, and represented all can impact the quality and use of the data. If the data is collected so that the validity and reproducibility is questionable, then that can impact data quality. The way in which the data are stored and processed can also impact quality. If we query a database and discover that the data are measured in different units for different patients, this can be an issue. For example, if weight was measured in kilograms for some patients and pounds for others, and if units were not stored in the database, than we have a huge data quality issue.



Whenever you're working with data, you must assume that errors exist. It is the job of the analyst to cleanse the data so that even though errors exist, the results of an analysis would not change. An experienced data analyst and health information management professional has the content knowledge and expertise to determine the best methods, given the study and context, to achieve valuable and valid conclusions with data, regardless of known errors.

Plan for Errors

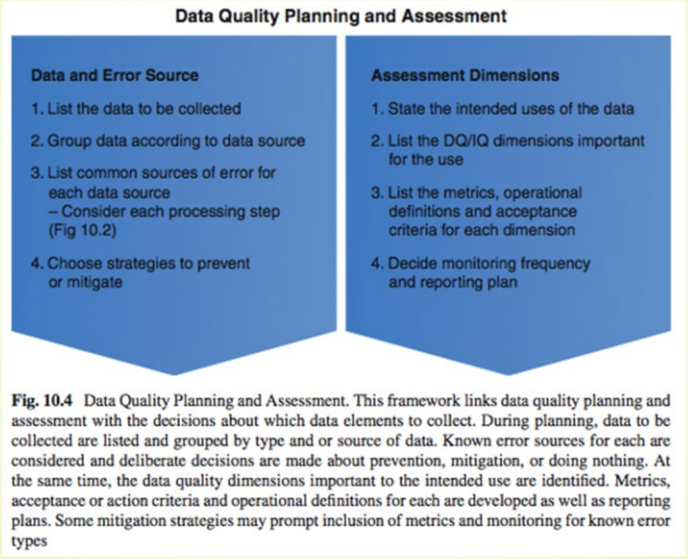
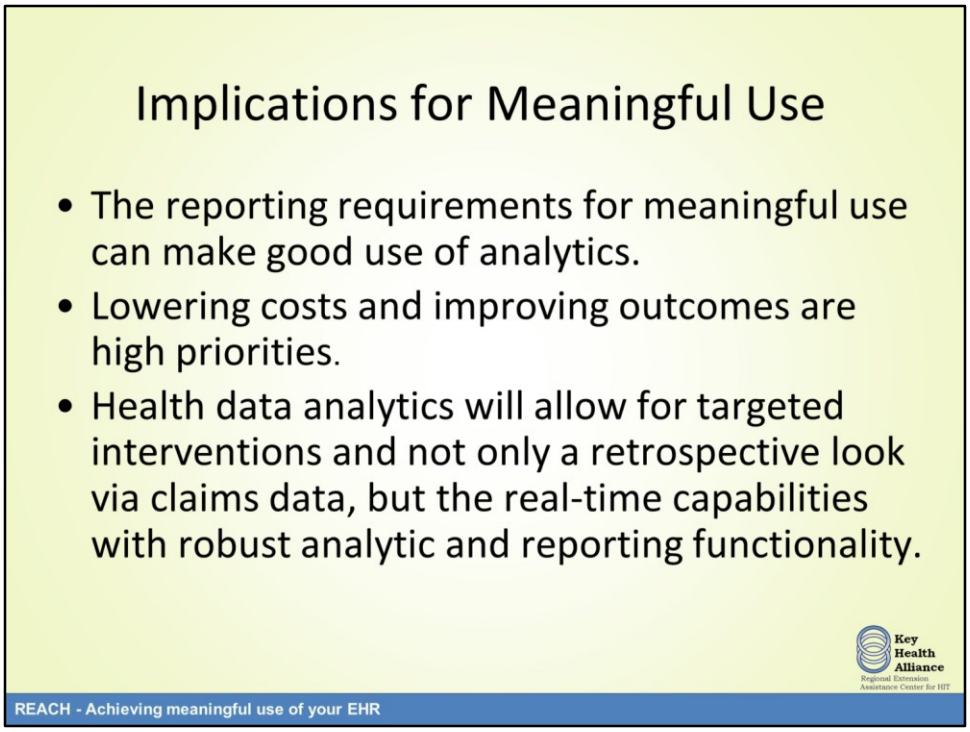


Fig. 10.4 Data Quality Planning and Assessment. This framework links data quality planning and assessment with the decisions about which data elements to collect. During planning, data to be collected are listed and grouped by type and or source of data. Known error sources for each are considered and deliberate decisions are made about prevention, mitigation, or doing nothing. At the same time, the data quality dimensions important to the intended use are identified. Metrics, acceptance or action criteria and operational definitions for each are developed as well as reporting plans. Some mitigation strategies may prompt inclusion of metrics and monitoring for known error types

Key Health Alliance
Regional Extension Assistance Center for HIT

It is always a good idea to plan for errors. The first thing to consider is the data that you intend to collect, its type and source. Also, list any known sources of error for each of these data types and sources, and choose the most appropriate strategy for dealing with the errors. You may have to develop a plan for error prevention or mitigation, or simply do nothing. At the same time you are dealing with the data and any errors, the data quality dimensions should be identified so that future issues with the data can be resolved.



Basic analytic methods have great implications for meaningful use. The reporting requirements for meaningful use can make good use of the analytic methods. The core and menu objectives require analysts to report proportions, therefore statistical techniques are quite minimal. However, lowering costs and improving outcomes is a priority of meaningful use. The goal is to move beyond the simple analytics required of meaningful use and examine the real value of having data readily available electronically. The adoption of health data analytics will allow for targeted interventions and not only a retrospective look of the data, but real-time capabilities with robust analytic and reporting functionalities.

References

Richesson, R. L., & Andrews, J. E. (Eds.). (2012). *Clinical research informatics*. Springer.

Key Health Alliance
Regional Extension Assistance Center for HIT