Journal of Nuclear Medicine, published on February 19, 2021 as doi:10.2967/jnumed.120.245654

Fundamental Statistical Concepts in Clinical Trials and Diagnostic Testing

Stephanie L. Pugh[1], Pedro A. Torres-Saavedra[1]

[1]NRG Oncology Statistics and Data Management Center, American College of Radiology, Philadelphia, PA

Corresponding Author:
Stephanie L. Pugh
pughs@nrgoncology.org
215-717-0850

Keywords: Hypothesis testing, multiplicity, diagnostic testing, receiver operating characteristic curves

Funding acknowledgement: U10CA180822 from the National Cancer Institute.

No potential conflicts of interest relevant to this article exist.

Running title: Introduction to Statistical Concepts

ABSTRACT

This article explores basic statistical concepts of clinical trial design and diagnostic testing. How one starts with a question, formulates it into a hypothesis upon which a clinical trial is then built, is integrated with statistics and probability, such as determining the probability of rejecting the null hypothesis when its actually true (type I error) and the probability of failing to reject the null hypothesis when the null hypothesis is false (type II error). There are a variety of tests for different types of data and the appropriate test must be chosen for which the sample data meet the assumptions. Correcting of the type I error in the presence of multiple testing is needed to control the error's inflation. Within diagnostic testing, identifying false positive and false negative patients is critical to understanding the performance of a test. These are utilized to determine the sensitivity and specificity of a test along with the test's negative predictive value and positive predictive value. These quantities, specifically sensitivity and specificity, are used to determine the accuracy of a diagnostic test using receiver operating characteristic curves. These concepts are briefly introduced, with references to allow the reader to explore various concepts at a more detailed level if desired, to provide a basic understanding of clinical trial design and analysis.

INTRODUCTION

Clinical trials and statistics serve as the basis of scientific research in biomedical sciences. Understanding the concepts is very important for the clinicians, investigators, and scientists working with statisticians on clinical trials. This paper focuses on basic statistical concepts, such as hypothesis testing, confidence intervals, parametric vs. non-parametric tests, multiplicity, and diagnostic testing, that form the building blocks of research. The NRG-HN006 trial in head and neck cancer, conducted by NRG Oncology, a research group funded by the National Cancer Institute, will serve as many of the examples for the statistical concepts presented.

NRG-HN006 TRIAL

There is a lack of consensus in the head and neck cancer community on how to treat patients with early stage oral cancer (*1,2*). NRG-HN006 randomizes T1-2N0M0 oral cavity patients with negative ^{18}fluorodeoxyglucose (FDG)-positron emission tomography (PET/CT) to elective neck dissection (END) or sentinel lymph node biopsy (SLNB) (NCT04333537). The co-primary objectives assess non-inferiority in disease-free survival (DFS) and superiority in quality of life (QOL).

For the SLNB arm, the primary tumor is injected with a radiotracer that travels to the cervical lymph nodes. The first echelon of nodes that are localized by the radiotracer represent the lymph nodes most likely to harbor metastatic disease. The SLN(s) can then be biopsied when the primary tumor is excised. Typically, a smaller incision(s) is made in the neck and less surgical manipulation is required to remove a small number of lymph nodes rather than to dissect the entire lymph node basin, as with END. Pathological examination is then focused on nodes with the highest likelihood to harbor disease, rather than on many nodes harvested from END. An important research question is whether there is a significant difference in the

performance of radiotracers in terms of the false-negative rate (FNR). The FNR, to be described later, corresponds to a negative SLNB result that develops subsequent metastatic lymph nodes without recurrence at the primary site (*3*).

HYPOTHESIS TESTING

Clinical trials are designed around a hypothesis that is used to determine its primary objective. Trials are conducted within a sample, a subset of the population of interest. *Statistics* are used to summarize the sample and estimate an unknown population *parameter*, a number summarizing the population (Table 1) (*4*). Hypothesis tests are based on a null hypothesis, H_0, and an alternative hypothesis, H_A. The *null hypothesis*, which is the hypothesis being tested, is a very specific statement about a parameter of the population. A broader statement that pairs with H_0, although mutually exclusive from it, is called the *alternative hypothesis*. This alternative hypothesis is sometimes referred as the *research hypothesis* as it states, in statistical terms using parameters, the primary hypothesis of the trial. For example, if the FNRs were compared between two radiotracers in NRG-HN006, *Rad1* and *Rad2*, H_0 and H_A would be:

$$H_0: FNR_{Rad1} = FNR_{Rad2} \ \ vs. \ \ H_A: FNR_{Rad1} \neq FNR_{Rad2}.$$

Hypothesis testing involving a symmetric alternative hypothesis like the one above would use a two-sided test. For a one-sided test:

$$H_0: FNR_{Rad1} \leq FNR_{Rad2} \ \ vs. \ \ H_A: FNR_{Rad1} > FNR_{Rad2.}$$

A trial with this H_A hypothesizes that radiotracer 1 has a higher FNR, and worse performance, compared to radiotracer 2 in the target population. Whether the test is one- or –two-sided is dependent on the question of interest, such as a primary or secondary objective, and is determined a priori.

Hypothesis testing is usually performed using a *test statistic*, which summarizes the sample information. Under a certain set of assumptions, a test statistic follows an exact or

approximate distribution under H_0 that reflects the randomness associated with the sample. The *p-value,* the probability of obtaining a statistic at least as extreme as the test statistic in the direction of H_A if H_0 were true, is used to interpret that test statistic[4]. The smaller the p-value, the stronger the evidence against H_0, leading one to reject it. Typically, this result is stated as being "statistically significant" in favor of H_A. Conversely, large p-values do not provide enough evidence against H_0 leading one to fail to reject it. Not being able to reject H_0does not make it true but rather conclude that there is not enough evidence to reject it.

Consider the two-sided test statistic for comparing the FNR between two radiotracers for SLNB in NRG-HN006. Previous studies have suggested that the FNR of the SLNB procedure can be around 5-15% (*5*). The value of the test statistic *Z* for comparing the FNR between radiotracer 1 (15%) and 2 (7%) observed with 154 patients per group, assuming a normal approximation, is 2.26 (Table 2) (*5*). *Z* is used to determine the p-value by matching this value to probabilities of the standard normal distribution. With a two-sided test, the p-value corresponding to z=2.26 is p=0.0238. The threshold, set a priori, to determine if the p-value is small enough to reject H_0 or large enough to fail to reject H_0 is known as the *significance level*. If the significance level=0.05, which is commonly used, then there is enough evidence to conclude that the FNR between the two radiotracers are different since p=0.0238 < 0.05 (i.e., H_0 is rejected). Statistical significance, however, does not provide evidence on the magnitude of the effect, making a statistically significant difference not necessarily clinically meaningful. For example, in a large sample size, a small effect can reach statistical significance due to the small variation in the sample. Likewise, large effects may fail to be deemed statistically significant if the sample is too small due to the large amount of chance variation (i.e. the analysis is underpowered).

The significance level also represents the probability of *type I error*, denoted as α. This error occurs when H_0 is rejected but it is actually true (Table 3). Thus, there is the truth for the

population and a decision to be made using the sample, yielding four possible scenarios. In addition to the type I error, the *type II error* is an incorrect decision that occurs when H_0 fails to be rejected but H_0 is actually false (i.e. H_A holds); its probability denoted as β. The two correct decisions are rejecting H_0 when it is false and failing to reject H_0 when it is correct.

Statistical power is related to the type II error by being its complement, 1-β. Thus, the statistical power of a hypothesis test is its ability to identify a specified effect size at α significance level; or, conversely, reject H_0 when H_A is true – a correct decision described above. Ideally, a trial should have large power to correctly conclude H_A when it is true. With continuous outcomes, four main components impact power: the specified effect size, the significance level, the sample size n, and the population variance σ^2. Specifically, power increases with larger effect sizes, higher α, larger sample sizes, and less variability within the sample. Most clinical trials are designed with statistical power ranging from 80% to 95%. Trials with power <80% or an overly optimistic hypothesized treatment effect size are usually considered underpowered (*6*).

Confidence intervals (CI) are used to determine the range of possible values of the true parameter, determined from the sample data, based on a certain level of confidence. For instance, 95% CIs are commonly used and indicate that with 95% confidence, the true value being estimated is within the interval. The level of confidence is determined by 1-α and, in general, is thus equivalent to the probability of failing to reject H_0 when H_0 is true. In many cases, since the level of confidence is determined based on the significance level, α, interpretation of the CI will correspond with that of the statistical test. For instance, if the FNR estimate for radiotracers 1 and 2 based on 154 patients per group is 15% and 7%, the 95% CI based on a normal approximation of the difference in FNR between the two radiotracers is (1.1%, 14.9%) (Table 2). The CI for the difference in FNR between the radiotracers does not contain 0, which would conclude that the radiotracers have different performance in terms of

FNR. This result corresponds to the p-value for the test in the prior example: $p=0.0238 < \alpha=0.05$, which produces a statistically significant result.

PARAMETRIC VS. NON-PARAMETRIC TESTS

When the assumption that the sample data follows a known probability distribution is met, such as the normal distribution, parametric tests can be used. The t-test, which is used to test the difference between two means is a parametric test that assumes the sample data comes from a normally distributed population (*7*). In large samples (e.g. >30) that do not meet the normality assumption, methods based on the normal distribution can still be used after invoking the *central limit theorem (CLT)*. Broadly speaking, the CLT states that regardless of the distribution of the population, as the sample gets larger, the distribution of the sample means approaches a normal distribution (*8*). This allows tests that assume data is normally distributed to be used to compare means. Versions of the t-test can be used in 2 independent samples or in paired samples (i.e. pre- and post-test). An *analysis of variance* is an extension of the t-test to more than two independent samples.

In small samples or those that draw from populations with heavily skewed distributions, non-parametric tests can be used instead. The distribution of the non-parametric test statistic can be derived under H_0 without specifying the underlying distribution of the population (*8*). The *Wilcoxon-Mann Whitney test* is the non-parametric version of the 2 independent sample t-test while the *Wilcoxon signed rank test* is the counterpart to the paired t-test (Table 4) (*9,10*). The Kruskal-Wallis test can be used to test differences between more than 2 independent groups. Non-parametric tests are not testing means, as in a t-test, but rather assign ranks to the data in order to test for differences in the groups' probability distributions and thus, typically report medians.

The *Chi-square* goodness-of-fit test was used when comparing FNR, a proportion, between two independent groups. Chi-square tests can be used for a single proportion, such as comparing FNR of a diagnostic test to a fixed value, or for two independent groups, as previously presented. It is a non-parametric test, since it does not require that the sample data follow a distribution, that uses frequencies from categorical or count data to describe how well this data fits with H_0. The expected value of 80% of the counts are required to be at least five for the test to have a good approximation to the chi-square distribution (*11*). If this assumption is violated, other tests, such as Fisher's exact test, can be considered. The *exact binomial test*, a parametric test based on the binomial distribution, can be used for binary data for a single proportion.

MULTIPLICITY

Recall that the type I error, α, is the probability of incorrectly rejecting H_0. In a single study with α=0.05, a type I error is expected to occur 5% of the time. Take the context of brain imaging with tests performed on each vertex of the image representation of the brain as an example (*12,13*). Roughly 100,000 voxels are obtained from a series of three-dimensional brain volumes with the same number of hypothesis tests to depict activated regions (*13*). If α=0.05, then 5,000 false positive (FP) results would be expected. Control of the *family-wise error rate (FWER)*, the probability of at least one type I error in the trial, is thus desired under the presence of multiple testing (*14*).

Multiple methods exist to control the type I error rate. The *Bonferroni correction* may be the most commonly used but it is also the most conservative, which can be desirable if strict control of the type I error is desired (*15*). When designing a study with co-primary endpoints, such

as NRG-HN006, a Bonferroni correction would require splitting the type I error. To maintain an overall α=0.05, each endpoint may use α=0.025. This increases the sample size required.

A study can be designed as to avoid the issue of multiplicity. *Hierarchical testing* is a method to control the type I error rate without affecting a clinical trial's sample size (*16*). In NRG-HN006, the co-primary endpoint of DFS is assessed first and if non-inferiority is shown, QOL superiority is tested. This allows both to use α=0.05 while maintaining an overall α=0.05.

Alternatively, control of the *false discovery rate (FDR)*, which is the proportion of significant results that are actually FPs, can be used to correct for multiplicity (*17*). FDR-based methods are often preferred at early stages of discovery due to their higher power to detect true positives while controlling the proportion of type I errors. A less conservative approach than the Bonferroni correction that is commonly used in function magnetic resonance imaging analysis is Hochberg's step-down procedure that adjusts for multiplicity by controlling the FDR (*13,17,18*). This procedure orders the p-values beginning with the least significant and compares each to an adjusted type I error, α'. Once $p < \alpha'$, then the comparisons stop and that test and all following tests are deemed statistically significant. Details on and comparisons of the corrections addressed here as well as additional ones, such as parametric tests, can be found elsewhere (*12-14,19*).

DIAGNOSTIC TESTING

In biomedical studies, diagnostic tests or procedures are typically used to determine the presence or absence of a disease or health condition. Diagnostic tests can be used for screening or surveillance, treatment monitoring, or staging. Some examples of diagnostic imaging tests are X-ray, PET, CT, PET/CT, MRI, and ultrasound (*20*). Test accuracy studies are usually designed to answer diagnostic or prognostic questions. Diagnostic test accuracy studies use the test information to classify a patient into a current health status while prognostic test

accuracy studies refer to the risk of a future health status. An example of a prognostic test accuracy study is given by the NRG-HN002 sub-study that estimated the accuracy of the 12-14 week post therapy FDG-PET/CT to predict 2-year loco-regional control (*21*). In general, a diagnostic test under study is also known as the "index test" (*22*). The true disease state is determined using a "gold standard" or "reference standard" test. In accuracy test studies with a diagnostic goal, index tests are usually proposed because they are associated with lower costs, faster results, or are less invasive. For instance, serology tests to detect the presence of antibodies in the blood when the body is responding to COVID-19 are considered index tests. These tests show if a person has been infected by coronavirus in the past. Antigen tests can also be considered index tests, but they instead diagnose active coronavirus infections. Antigen tests have a higher chance of missing an active infection, so negative test results are usually confirmed with a molecular test. Due to their high diagnostic accuracy, molecular tests such as the nucleic acid amplification test are considered the gold standard tests to determine if a patient has COVID-19. Several antigen and antibody tests have been proposed due to their lower costs and sometimes faster results. In the NRG-HN002 prognostic test accuracy sub-study, the FDG-PET/CT at 12-14 weeks post-treatment is the index test and the protocol-specified methods to assess loco-regional failure at 2 years post-randomization is the reference standard (*21*).

Continuing with the NRG-HN006 example of radiotracers, the SLNB with a given radiotracer is the index test, which was used to determine lymph node metastasis. The SLNB result is a positive or negative nodal metastasis according to the pathology findings from the SLNB. The subsequent development of isolated cervical metastasis assessed through standard imaging following the SLNB is the "reference standard." A patient is called a *false negative (FN)* if she/he has lymph node metastases but the SLNB gives a negative result (Table 5). Conversely, a patient is called a *FP* if she/he does not have lymph node metastases but the SLNB predicts a

positive result. The *FNR*, a measure to assess the performance of a diagnostic test, determines the proportion of incorrect negative test results among individuals with the disease. *Sensitivity*, 1-FNR, or *true positive rate*, of the test indicates the probability of a positive result among those with the disease (D^+) (Table 6) (*23*). Similarly, the *False Positive Rate (FPR)* determines the proportion of incorrect positive test results among those without the disease (D^-). The *specificity* of the test, 1-FPR, or *true negative rate*, indicates the probability of a negative result among those individuals without the disease. The ideal diagnostic test should have high specificity and sensitivity (*24*). A trade-off between specificity and sensitivity depends on whether the diagnostic test is used for screening, staging, or prognosis.

In SLNB, an objective can be to estimate the ability of the SLNB to predict a N0 neck result (i.e., no lymph node metastasis) since these patients may avoid an unnecessary neck dissection. That is, what is the probability of developing isolated cervical metastasis after a negative SLNB (i.e. N0 neck)? The *Negative Predictive Value (NPV)* of a test indicates the probability of not having the disease given a negative test result. Likewise, the *Positive Predictive Value (PPV)* represents the probability of having the disease given a positive test result. The complements of the NPV and PPV are called the *false omission rate (FOR)* and *FDR*, respectively. While the sensitivity and specificity are quantities inherent to the performance of the diagnostic test, the NPV and PPV depend not only on the test's performance but also on the prevalence of the disease or health condition (Figure 1).

In the SENT trial, patients with negative SLNB who subsequently developed cervical metastasis and had a negative primary tumor site were classified as FNs (*25*). It is typical in SLNB studies that the number of FPs is deliberately kept zero since a positive SLNB result is deemed sufficient to declare the presence of cervical nodal metastases (Table 7) (*26*). That is, the specificity and PPV of the SLNB are both 100% (FPR is 0%). Occult lymph node metastases not detected by the SLNB (FNs) is of concern to clinicians since these patients may

receive alternative therapies such as close observation for low-risk patients (*2*). Patients with occult nodal metastasis may be at risk of distant metastatic disease given that the cancer has spread out to the lymph node basins. However, if the SLNB predicts N0 necks with high probability, these patients may avoid unnecessary therapy and its implications relative to morbidity, decreased QOL, and cost. FNs in SLNB can occur because the lymphatic pathway to the involved node is blocked, the pathologist fails to detect micrometastasis or isolated tumor cells inside a lymph node, or the surgeon misses a positive sentinel lymph node due to poor training or complexity of the surgical region (*26*). An estimate of the FNR for SLNB in oral cancer is 15/109=0.138 (13.8%) (Table 7). Assuming normality, a 95% CI for the FNR is (0.073, 0.203) which indicates that the true FNR is between 7.3% and 20.3% with 95% confidence. This FNR estimate for the SLNB is of concern to some clinicians since roughly 1 or 2 out of 10 patients could be incorrectly diagnosed. The NPV for the SLNB to detect N0 neck patients is given by 306/321=0.95. Similarly, the sensitivity and specificity of the SLNB is 0.86 (94/109) and 1.00 (306/306), respectively. For a given patient, the probability of having lymph node metastasis after a negative SLNB result increases to 0.95 from 0.74, the latter being the probability of no nodal metastasis before the SLNB. It is important to interpret the NPV (and PPV) after considering disease prevalence. For a given sensitivity and specificity rate, the NPV increases as the prevalence of the disease, $P(D^{+})$, decreases (Figure 1). See Civantos et al and Hines et al as additional examples of these terms (*26,27*).

An example of a prognostic test accuracy study is a potential NRG-HN006 sub-study assessing the predictive accuracy (NPV) of the FDG-PET/CT when combined with the END or SLNB to predict 1-year loco-regional control. So, patients with negative FDG-PET/CT and negative END or SLNB result (“index test”) would have loco-regional control assessed at 1 year using standard imaging and a biopsy confirmation (“reference test”) per protocol-specified techniques. Note that only the row with the negative index test results from Table 5 is included

in the study by design. Given the negative index results, the loco-regional control rates at 1 year can be compared using a Chi-square test.

When designing a test accuracy study, it is crucial to carefully examine the objectives and, therefore, the design type as it dictates what accuracy measures can be properly estimated from the data. For instance, the NPV and PPV cannot be estimated from a case-control design given that the proportion of patients with the disease based on the reference standard is manipulated by researchers, for example, by setting a 1:1 case-control matching (*28*). One of the NRG-HN006 eligibility criteria is an FDG-PET/CT negative result for lymph node metastasis. Thus, a reasonable inference target would be to estimate the NPV of FDG-PET/CT in this population within the END arm only since the number of patients with a negative index test is fixed by researchers through the trial design. In this case, the true metastatic nodal status *D* is determined by the pathological findings after the END.

ROC ANALYSIS

In many applications, investigators use continuous or ordinal biomarkers, or build predictive models based on a continuous or ordinal scale using a combination of variables such as biomarkers, gene expressions, and patient's characteristics, among others (*29*). A single biomarker or predictive model can be regarded as a classifier for purposes of diagnostic testing. These classifiers can, however, be converted into a binary classifier after selecting a given threshold on a suitable scale. For instance, logistic regression models are usually employed to construct classifiers based on a set of predictors (*30*). Often, thresholds are selected on a probability scale. For instance, if a patient has a predictive probability based on the logistic model >0.5, then that patient will be considered a positive result for diagnostic purposes. This binary classifier based on a threshold can be then framed within the binary diagnostic testing

discussion presented in Table 5. The selection of a threshold should follow some type of optimality criterion to obtain a classifier ("diagnostic test") with at least acceptable accuracy. The discriminative power or diagnostic performance of a classifier is usually summarized and measured using the area under the curve (AUC) of the *receiver operating characteristic (ROC) curve* (*31*). The ROC curve plots the FPR (1-specificity) by sensitivity for different thresholds (Figure 2). A classifier that perfectly predicts the disease status among those with and without the disease has an AUC=1. Randomly predicting the disease status leads to a classifier with an AUC=0.5. The AUC can be also be interpreted using probabilities. Assume a rater is asked to score two individuals, one with the disease and other without the disease. The AUC can be seen as the probability that the rater will give the individual with the disease a higher score than that without the disease. An alternative interpretation of the AUC is the average sensitivity across all possible FPRs. See Hyun et al for an example utilizing AUC (*32*).

The goal in a ROC analysis is, therefore, to select an optimum threshold that produces a classifier closer to the upper left corner of the graph. Note that for a random classifier (i.e. classification of an individual within each disease status is done randomly with equal probability, using, for instance, a fair coin) the NPV=1-P(D^+). This result tells us that the classifier does not improve the predictive ability of non-disease.

The ROC AUC is a statistic allowing typical inferential procedures to be applied. Namely, it is possible to perform hypothesis testing and CI estimation for the AUC. Likewise, it is possible to compare the AUC for two or more groups.

CONCLUSIONS

Clinical trials, the gold standard in research, are based on various statistical concepts and assumptions. The probability of type I and type II errors are specified in advance and

impact the rigor of the study's conclusions. The number of hypothesis tests being conducted can inflate the type I error resulting in the necessity to control the FWER. When performing diagnostic testing, one must be aware of various performance measures such as sensitivity and specificity, which are used to create a ROC curve that depicts the discriminative power of a diagnostic test or classifier. Having a basic understanding of these concepts can aid an investigator interested in conducting research and understanding how the results inform the conclusion of research publications.

References:

1. de Bree R, Takes RP, Shah JP, et al. Elective neck dissection in oral squamous cell carcinoma: Past, present and future. Oral Oncol. 2019;90:87–93.
2. Hutchison IL, Ridout, F, Cheung SMY, et al. Nationwide randomised trial evaluating elective neck dissection for early stage oral cancer (SEND study) with meta-analysis and concurrent real-world cohort. Br J Cancer. 2019;121:827–836.
3. Schilling C, Stoeckli SJ, Vigili MG, et al. Surgical consensus guidelines on sentinel node biopsy in patients with oral cancer. Head Neck. 2019; 41:2655-2664.
4. Pugh SL, Molinaro A. The nuts and bolts of hypothesis testing. Neurooncol Pract. 2016;3:139-144.
5. Schilling C, Stoeckli SJ, Haerle SK, et al. Sentinel European Node Trial (SENT): 3-year results of sentinel node biopsy in oral cancer. Eur J Cancer. 2015;51:2777-84.
6. Laino, Charlene. Study: Many ASCO Meeting Phase III Trials Underpowered/ Oncology Times. 2007;29:58-59.
7. Altman DG, Bland MJ. Parametric *v* non-parametric methods for data analysis. BMJ. 2009;338:a3167.
8. Conover WJ. Probability Theory and Statistical Inference. In Conover WJ, Practical nonparametric statistics. Third Edition. Hoboken, NJ: Wiley 1999.
9. Forrester JC, Ury HK. The Signed-Rank (Wilcoxon) test in the rapid analysis of biological data. Lancet. 1969;1:239-241.
10. Divine G, Norton HJ, Hunt R, Dienemann J. Statistical grand rounds: a review of analysis and sample size calculation considerations for Wilcoxon tests. Anesth Analg. 2013;117:699-710.
11. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb). 2013;23:143-149.

12. Lindquist MA, Mejia A. Zen and the art of multiple comparisons. Psychosom Med. 2015;77(2):114-125.
13. Alberton BAV, Nichols TE, Gamba HR, Winkler AM. Multiple testing correction over contrasts for brain imaging. Neuroimage. 2020;216:116760.
14. Dmitrienko A, Tamhane A.C. and Bretz, F. Multiple Testing Problems in Pharmaceutical Statistics. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series; 33; 2010.
15. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34:502-508.
16. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group sequential trials. Stat Med. 2010;29:219-28.
17. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res. 2008;17(4):347-388.
18. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75:800-802.
19. Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis. 2017;9:1725-1729.
20. American Society of Clinical Oncology, Tests and Procedures (March, 24 2020). https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures. Accessed July 6, 2020.
21. Subramaniam RM, Demora L, Yao M, et al. 18 FDG PET/CT prediction of treatment outcomes in patients with p16-positive, non-smoking associated, locoregionally advanced oropharyngeal cancer (LA-OPC) receiving deintensified therapy: Results from NRG-HN002. JCO. 2020;38(15_suppl):6563-6563.

22. Bossuyt PM, Reitsma JB, Bruns DE, et alSTARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies, https://www.equator-network.org/reporting-guidelines/stard/. Accessed July 10, 2020.
23. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ.* 1994;308:1552.
24. Glasser SP. Research Methodology for Studies of Diagnostic Tests. In: Glasser SP (eds) Essentials of Clinical Research. Dordrecht: Springer; 2008:245-257.
25. Kataria K, Srivastava A, Qaiser D. What Is a False Negative Sentinel Node Biopsy: Definition, Reasons and Ways to Minimize It? Indian J Surg. 2016;78:396–401.
26. Civantos FJ, Zitsch RP, Schuller DE, et al. Sentinel lymph node biopsy accurately stages the regional lymph nodes for T1-T2 oral squamous cell carcinomas: results of a prospective multi-institutional trial. JCO. 2010;28:1395-1400.
27. Hines JP, Howard BE, Hoxworth JM, Lal D. Positive and Negative Predictive Value of PET-CT in Skull Base Lesions: Case Series and Systematic Literature Review. J Neuro Surg Rep. 2016;77:e39–e45.
28. Mathes T, Pieper D. An algorithm for the classification of study designs to assess diagnostic, prognostic and predictive test accuracy in systematic reviews. Syst Rev. 2019;8:226.
29. Baker SG. The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer. JNCI. 2003;95:511–515.
30. Agresti A. Logistic Regression. In:Agresti A, Categorical Data Analysis. Wiley Series in Probability and Statistics, Second Edition, Hoboken, NJ; 2002.
31. Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. Acta Pædiatrica. 2007; 96:644-647.
32. Hyun OJ, Luber BS, Leal JP, et al. Response to Early Treatment Evaluated with 18F-FDG PET and PERCIST 1.0 Predicts Survival in Patients with Ewing Sarcoma Family of

Tumors Treated with a Monoclonal Antibody to the Insulinlike Growth Factor 1 Receptor. J Nucl Med. 2016;57:735-740.

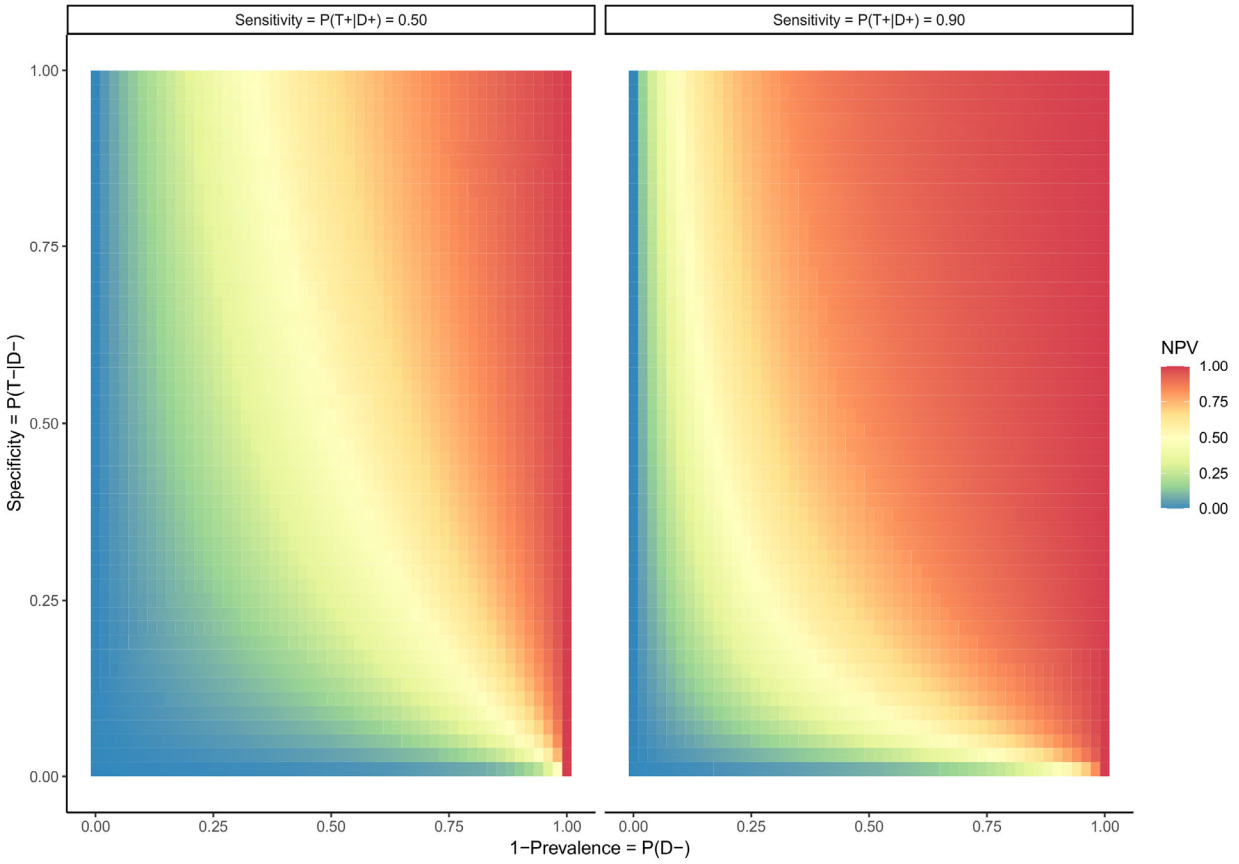

Figure 1. The relationship between negative predictive value, specificity, and disease prevalence when sensitivity of a diagnostic test is 50% and 90%.

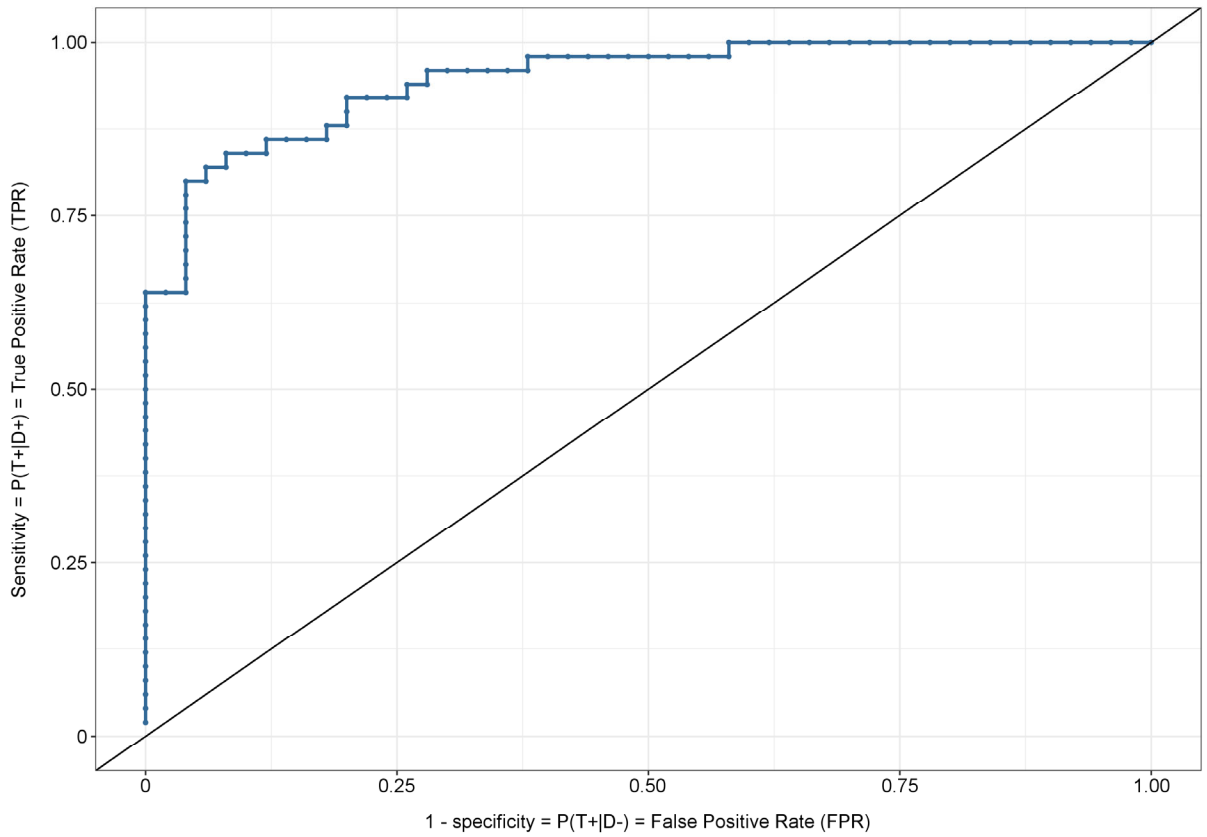

Figure 2. Receiving Operating Characteristic (ROC) Curve. Each point along the ROC curve represents a set of coordinates (1-specificity, sensitivity) for a classifier defined by a threshold. The diagonal line represents a random classifier.

Table 1. Statistical Terms

Term	Definition	Brief Example
Statistic	Summarizes the sample and estimates an unknown population parameter	NPV estimated from a sample
Parameter	Number summarizing the population	NPV of a test in the population
Null hypothesis (H_0)	Specific statement about parameters of the population	$H_0: NPV_{PET/CT} < 90\%$
Alternative hypothesis (H_A)	Broad statement that pairs with, yet is mutually exclusive from, H_0	$H_A: NPV_{PET/CT} \geq 90\%$
Test statistic	Summarizes the information from the sample	When comparing two means assuming a normal distribution, Z is the test statistic, Z, follows standard normal distribution
p-value	Probability of obtaining a sample statistic at least as extreme than the test statistic in the direction of H_A if H_0 were true	Z=2.26, calculated from comparing 154 patients with an observed FNR=15% to 154 patients with an observed FNR=7%, corresponds to p=0.0238.
Type I error (α)	Probability of rejecting H_0 when true	Phase 3 superiority trials are commonly designed with a 1-sided type I error=0.025.
Type II error (β)	Probability of failing to reject H_0 when false (i.e. H_A holds).	When designing a clinical trial, the type II error is set a priori with β=0.05-0.20 commonly used.
Statistical power (1-β)	Probability of rejecting H_0 when H_A is true	Clinical trials are commonly designed with 80-95% power.
Confidence interval (CI)	Provides a range of possible values of the true parameter based on a specified level of confidence	Pathologic analysis of SLNs by routine hematoxylin and eosin revealed NPV=0.94, 95% CI: 0.88-0.98 (*26*).
Family-wise error rate control	Control of the probability of at least one type I error	Bonferroni correction divides the type I error by the number of tests.
False discovery rate control	Control of the proportion of significant results that are actually false positives	Hochberg's step-down procedure orders p-values to compare to an adjusted α

Table 2. Equations for Comparing the False Negative Rate (FNR) of Two Radiotracers

Test statistic Z for a comparison of two binomial samples using the normal approximation	$$Z = \frac{\widehat{FNR}_{Rad1} - \widehat{FNR}_{Rad2}}{\sqrt{\frac{\widehat{FNR}_{Rad1} * \left(1 - \widehat{FNR}_{Rad2}\right)}{n_{Rad1}} + \frac{\widehat{FNR}_{Rad2} * \left(1 - \widehat{FNR}_{Rad2}\right)}{n_{Rad2}}}}$$ $$= \frac{0.15 - 0.07}{\sqrt{\frac{0.15 * 0.85}{154} + \frac{0.07 * 0.93}{154}}} = 2.26$$ where $\widehat{FNR}$=FNR estimated from the sample to estimate the population FNR and n=number of patients receiving each radiotracer
95% confidence interval for the difference of two proportions using the normal approximation	$$\left(\widehat{FNR}_{Rad1} - \widehat{FNR}_{Rad2}\right)$$ $$\pm z\sqrt{\frac{\widehat{FNR}_{Rad1} * \left(1 - \widehat{FNR}_{Rad2}\right)}{n_{Rad1}} + \frac{\widehat{FNR}_{Rad2} * \left(1 - \widehat{FNR}_{Rad2}\right)}{n_{Rad2}}}$$ $$= 0.08 \pm 0.069 \rightarrow [1.1\%, 14.9\%]$$ where z corresponds to a quantile of the standard normal distribution for the chosen confidence level, 95%.

Table 3. Probabilities Associated with Hypothesis Testing.

	Truth	
Result of Statistical Test	**H_0 is true**	**H_0 is false (H_A holds)**
Fail to Reject H_0	Correct decision	Type II error (β)
Reject H_0	Type I error (α)	Correct decision (1- β)

H_0 represents the null hypothesis and H_A the alternative hypothesis

Table 4. Parametric vs. Non-parametric tests

	Parametric	**Non-Parametric**
Comparison of two independent groups with continuous outcomes	t-test	Wilcoxon-Mann Whitney test
Comparison of more than two independent groups with continuous outcomes	Analysis of variance	Kruskal-Wallis test
Comparison of two paired samples with continuous outcomes	Paired t-test	Wilcoxon signed rank test
Single proportion	Binomial Exact Test	Chi-square test

Table 5. SLNB (index test) result and pathology/neck dissection ("reference standard") result

	Isolated cervical metastases following SLNB (True disease state)		
SLNB result based on sentinel lymph nodes	Negative	Positive	**Total**
Negative	TN	FN	**T^-**
Positive	FP	TP	**T^+**
Total	**D^-**	**D^+**	**n**

TN=True Negative; TP=True Positive; FP=False Positive; TP=True Positive; D^-, D^+=number of patients without and with true nodal metastasis; T^-, T^+=number of patients with negative and positive SLNB results; n=total number of patients.

Table 6. Diagnostic Testing Terms

Term	Definition	Brief Example
False positive rate (FPR)	Proportion of incorrect positive test results among those without the disease	FPR of sentinel lymph node biopsy in T1-2 oral squamous cell carcinomas was 29.3% (*26*).
Specificity (1-FPR)	Probability of a negative result among those individuals without the disease (true negative rate)	Specificity of sentinel lymph node biopsy in T1-2 oral squamous cell carcinomas was 70.7% (*26*).
False negative rate (FNR)	Proportion of incorrect negative test results among individuals with the disease	FNR of sentinel lymph node biopsy in T1-2 oral squamous cell carcinomas was 9.8% (*26*).
Sensitivity (1-FNR)	Probability of a positive result among those individuals with the disease (true positive rate)	Sensitivity of sentinel lymph node biopsy in T1-2 oral squamous cell carcinomas was 90.2% (*26*).
Negative predictive value (NPV)	Probability of not having the disease given that the test result was negative	NPV in NRG-HN002 for 2-year loco-regional control of the head and neck was 94.5% (*24*).
Positive predictive value (PPV)	Probability of having the disease given that the test result was positive	PPV of radiologist's interpretation of skull base lesions, SUV cutoff of 2.5, and SUV cutoff of 3.0 was 80%, 60%, and 68.4%, with NPV=100%, 83.3%, and 75%, respectively (*27*).
Receiver operating characteristic (ROC) curve	Plot of a diagnostic tests' 1-specificity by sensitivity for different thresholds	Hyun et al (*32*).
Area under the curve (AUC) of ROC	Measure for how well a classifier can differentiate between two diagnostic groups	ROC AUC=0.71 when predicting 1-year overall survival from changes in (18)F-FDG uptake after therapy for Ewing sarcoma family of tumors (*32*).

FDG-PET/CT=[^{18}F]fluorodeoxyglucose-PET/CT; SUV=standardized uptake value.

Table 7. Results of SLNB in the SENT study (*3*)

	Isolated cervical metastases following a SLNB (Reference standard)		
SLNB result (Index test)	Negative (D^-)	Positive (D^+)	**Total**
Negative (T^-)	306	15	**321**
Positive (T^+)	0	94	**94**
Total	**306**	**109**	**415**

SLNB=Sentinel Lymph Node Biopsy