Official Title: An Intraoperative Guidance Platform for Radio Frequency Ablation
NCT number: NCT04152343
Document Type: Statistical Analysis Plan
Date of the Document: July 10, 2018

PRINCIPAL INVESTIGATOR:

Eric Hoffer, MD
Department of Radiology
Section of Interventional Radiology
Dartmouth-Hitchcock Medical Center
One Medical Center Drive
Lebanon, NH 03756
Tel: 603-650-7417
Email: Eric.K.Hoffer@hitchcock.org

As noted in the protocol, the following Statisical Analysis Plan is referenced in the protocol (page 9). The folllowing was obtained from the study SF19108 clinical protocol v29July2019, pages 13-18, Section 6.0 Statistical Plan

6 Statistical Plan

6.1 Population Size

Fifty two (52) patients will be recruited in the study, anticipating that up to 4 patients might drop out from the study, leaving a total of 48 patients. Based on prior statistics from Dr. Hoffer's clinic, approximately 30% of patients have more than one tumor which is treated. Assuming that 30% of the patients have two tumors,

recruitment of 48 subjects will yield 62 treated individual tumors which will be analyzed. A population of 62 analyzed tumors provides a sufficient statistical power for the outcomes the trial is designed to estimate, given reasonable hypotheses on the effect size that the PGP will result in, as discussed next.

6.2 Statistical Design

A single-arm trial of the research PGP computer technology is proposed in which a comparison of outcomes is made to benchmarks based on prior studies. There are two primary outcomes, technical success and clinical success. Because the measurement process is drastically different for these outcomes, the statistical analysis also differs markedly for each outcome and so each is separately described. Results from Kim, et al. [5] provide the benchmark against which the *technical* success of PGP is assessed. Results from Mulier, et al. [4] provide the benchmark against which the *clinical* success of PGP is assessed. We will recruit patients with a tumor size >2cm, and we will categorize tumor size as small if < 3cm, medium if between 3cm and 5cm, and large if > 5cm. In our analyses for clinical outcomes, we will weight our sample to equate to the relative frequencies of small, medium and large tumors to ensure that the new trial data is relevant to the benchmark from the historical data [4]. In addition to performing overall analyses and statistical tests, we will also report estimates broken down by tumor size and other subgroups of interest.

6.3 Statistical analysis of technical success

The ideal outcome, from a technical standpoint, is complete removal of the tumor and absence of cancer in an area of 3 mm beyond the region where the tumor appeared on the interventional contrast CT image (the "ablative margin"). In the past, **such an outcome was rare with only 3 successful cases out of 110** (see Kim, et al., 2010). However, it is notable that in previous work [5] the tumor was found to recur in 0/15, 2/34 (5.9%), 10/53 (18.9%), and 25/110 (22.7%) cases when the attained ablative margin was ≥ 3mm, ≥ 2mm, ≥ 1mm, and ≥ 0mm, respectively. Therefore, it is reasonable to expect that attainment of an ablative margin of ≥ 3mm would be a very successful outcome, and we will analyze this event as a secondary outcome.

The dependent variable in the primary analysis of technical success is the binary indicator of whether or not the RFA-GP surgery resulted in an ablative margin of ≥ 3 mm. Let p_{RFA} denote the true proportion of times an ablative margin of ≥ 3 mm is obtained in the proposed PGP study. As a benchmark against which to compare p_{RFA}, we deliberately err on the side of over-estimating the success of the standard procedure by using $p_{\mathrm{base}} = \frac{3}{110} + \frac{z_{0.975}}{110}\left(\frac{3(110-3)}{110}\right)^{1/2} = 0.0577$, where $z_{0.975}$ is the 97.5'th quantile of the standard normal distribution, as the benchmark. Note that p_{base} is the upper end-point of an approximate 95% confidence interval for the true proportion of times an ablative margin of ≥ 3 mm is obtained in the absence of PGP in [5]. We will test the one-sided hypothesis:

$$H_0: p_{\mathrm{RFA}} \leq p_{\mathrm{base}} \text{ against the alternative } H_1: p_{\mathrm{RFA}} > p_{\mathrm{base}}$$

using a one-sample test of proportions.

Because technical success will be assessed by each of three raters, the hypothesis test given above can be tested in a manner that allows for the effects of each rater and the clustering of ratings by subject to be accounted. Let TS_{ij} denote the technical success rating of subject *i* by rater *j*. We may then estimate the Rasch model

$$TS_{ij} \mid \theta_i \sim Bern(p_{ij})$$

where

$$p_{ij} = \Pr\left(TS_{ij} = 1 \middle| \theta_i\right) = \frac{\exp(\theta_i + \beta_j)}{1 + \exp(\theta_i + \beta_j)}$$

and $\theta_i \sim \mathrm{normal}(\theta_{RFA}, \sigma^2)$ describes the distribution of the log-odds of technical success across the population of subjects, accounting for the clustering of ratings by subjects, and $\beta_1 + \beta_2 + \beta_3 = 0$ is an identifiability constraint. Note that we use this identifiability constraint in lieu of treating $\beta_1, \beta_2, \beta_3$ as random-effect parameters, which is not advisable with only three raters. The target of inference is θ_{RFA}, the population mean log-odds of success, and the comparison value is

$$\theta_{base} = \log\left(\frac{p_{base}}{1 - p_{base}}\right)$$

Therefore, with data from multiple raters, we may transform the above hypothesis test to:

$$H_0: \theta_{\text{RFA}} \leq \theta_{\text{base}} \text{ against the alternative } H_1: \theta_{\text{RFA}} > \theta_{\text{base}}$$

or define

$$p_{RFA} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)}$$

and test $H_0: p_{\text{RFA}} \leq p_{\text{base}}$ against the alternative $H_1: p_{\text{RFA}} > p_{\text{base}}$. The former is preferred as the parameters are unrestricted. In addition to reporting p-values, we will also report the estimated probability for the proportion of occasions that an ablative margin of ≥ 3mm was obtained and an associated 95% confidence interval for p_{RFA} both in the full sample and for each tumor size subgroup.

We realize that the comparison to a fixed literature-based historical control value is non-ideal. Ideally, individual patient data from multiple historical studies performed in the absence of PGP that can be analyzed simultaneously with the data from the new study. This would allow us to account for study-to-study variability and to test the hypothesis that PGP results in a higher probability of technical success. If patient characteristics and baseline clinical information was available across the studies we could also adjust for such factors to make the studies more homogeneous. In lieu of having such data, we err on the side of conservatism by making the comparison point the upper limit of a 95% confidence interval based on the non-PGP study result, a threshold for success that is almost certainly more demanding than if we had the just described data and performed a statistical test for a two-population comparison. What we are essentially proposing to do is to reject the null hypothesis that the lower 95% confidence interval in the PGP study exceeds the upper 95% limit in the non-PGP study. Note that the non-overlap of two confidence intervals for two groups implies that a hypothesis test will reject the null hypothesis of no difference between the groups. Therefore, if we reject the null hypothesis in the proposal of the probability under PGP being less than or equal to θ_{base}, or to p_{base}, we can be very confident that PGP has a higher probability of technical success.

Finally, to evaluate the relationship between tumor size and the likelihood of obtaining an ablative margin ≥ 3mm, we will estimate a logistic regression model in which tumor size is a lone predictor. That is, we will estimate the value of $\beta = (\beta_0, \beta_{\text{size}})$ in the model:

$$\Pr(\text{Ablative margin} \geq 3\text{mm} \mid \text{size}) = \frac{\exp(\beta_0 + \beta_1 \text{size})}{1 + \exp(\beta_0 + \beta_1 \text{size})}$$

so that the probability of technical success when PGP is used is able to be predicted for a patient with a given exact tumor size. This expression will be helpful if it is necessary to determine the expected technical success for the size distribution of a sample of patients analyzed in a prior study. In the meantime, the fitted logistic regression model will only be used as a means of graphically summarizing the results with respect to tumor size.

6.3.1 Power for technical success

If the analysis of the above one-sided hypothesis test for p_{RFA} for an ablative margin of 3mm is conducted using Fisher's exact test, then with a sample-size of 62 lesions and allowing a type I error of 0.05, we can detect a value of $p_{\text{RFA}} > 0.161$ with 80% power. That is, the minimal detectable p_{RFA} difference is 0.103. Subgroup analyses (e.g., based on treated tumor size) will require larger effect sizes to retain the same level of power. However, the study is only intended to be powered for the primary analysis.

6.4 Statistical analysis of clinical success

It has been noted in the literature that if a tumor reoccurs in the same site, it typically does so within the first two years [4]. There is some evidence of reoccurrence as far out as 40 months but that is based on limited data and beyond 4 years there is no follow-up data [4]. For the purpose of this study, the analysis of the clinical outcome will be based on a 2-year endpoint. As with the determination of technical success,

a literature-based benchmark will be used in lieu of a concurrent control. Results in [5] suggest that the cumulative incidence of reoccurrence by 2 years is around 26.4% and by 4-years is around 37.8%. However, these results were not stratified by tumor size, which is known to be a very strong predictor of clinical success. In [4], reoccurrence rates of 14.1%, 24.5%, and 58.1% were reported for the small, medium and large tumor groups. However, these rates were evaluated over variable lengths of follow-up. The proportion of cases expected in the small, medium, and large categories are 12.9%, 41.9%, and 45.2% respectively (based on past clinical experience). The benchmark number to be used for the overall group comparison is the weighted average of these proportions with the above recurrence rates, which equals 24.1%. Because the failure rate data in [4] involve 1817 observations, we use this number as a basis for quantifying the precision in the baseline. Computing the lower end-point of a 95% confidence interval, we obtain $q_{\text{base}} = 0.2217$ as our ideal or most conservative benchmark against which to compare outcomes under PGP. This value is conservative even beyond the buffer provided by the allowance for uncertainty in its true value via the confidence interval end-point calculation as the follow-up times in [4] were shorter on average than 2-years.

For both the conservative and less conservative benchmarks, the overall assessment of the clinical success of PGP will be evaluated using the hypothesis test:

$$H_0: q_{\text{RFA}} \geq q_{\text{base}} \text{ against the alternative } H_1: q_{\text{RFA}} < q_{\text{base}}$$

where q_{RFA} is the probability of reoccurrence by 2-years of follow-up.

Patients are enrolled in the study over a 2-year period. Therefore, if every patient was followed up for 2-years, the study would be 4-years long. We propose a statistical analysis plan which will allow estimating the clinical performance of the PGP at any point after the initial recruitment phase of 2 years – with increasing accuracy as the analysis approaches the 4-year mark, having access to 2-year follow up data for each patient who entered the study.

The analysis will be as follows: the investigators will check the status of patients every 6-months (data may be available every 3-months for some patients but to be safe we only assume 6-month interval data will be available) and essentially impute future status from patients' prior status. The imputations will be automatically performed by a Bayesian analysis, which assigns an unknown parameter to each participant's 6-monthly future status, through 2-years of follow-up.

The data for each patient is a 5-item multinomial random variable indicating whether they experienced recurrence before 6-months, between 6 and 12 months, between 12 and 18 months, between 18 and 24 months, or survived recurrence-free to 24 months and a covariate for the size of tumor. The four values of the multinomial outcome variable are (1,0,0,0,0) for recurrence before 6 months, (0,1,0,0,0) for recurrence between 6 and 12 months, (0,0,1,0,0) for recurrence between 12 and 18 months, (0,0,0,1,0) for recurrence between 18 and 24 months, and (0,0,0,0,1) for no recurrence in the 24-months follow up period.

We refrain from using a Cox model or a parametric survival model to compare the status of the patients at two years of follow-up because reoccurrence status may only be known at 6-monthly intervals. Instead, we use a discrete-time survival model in which we model the probability of reoccurrence in each 6-month period thru 2-years of follow-up along with the event that the length of follow-up without reoccurrence exceeds 2-years. The dependent variable is a polytomous multinomial random variable because each patient's reoccurrence status (if followed up for the full 2 years) will be in exactly one of the 5 categories. The probabilities for the 5 categories sum to 1. The purpose of modeling the interim survival probabilities is that the interim probabilities allow the missing outcome status for patients not followed for two-years to be internally imputed during Bayesian model estimation. This approach allows for maximal use of the information in the data as all subjects are used in the analysis even if that haven't been followed up for two years (emulating what the Cox or other survival model does with continuous-time follow-up data). If we end up measuring reoccurrence status every three months, we will still pursue the described approach (but have 9 categories for the multinomial outcome as opposed to 5 categories) as the data will still not be close enough to continuous to support using the Cox model.

6.4.1 Notation for multinomial (discrete-valued) outcome

Let $y_{ij} = (y_{ij0}, y_{ij1}, y_{ij2}, y_{ij})$ denote the multinomial outcome for the *i*'th patient as assessed by the *j*'th rater ($j = 1,2$) and x_i denotes the tumor size for the i^{th} of $n = 62$ lesions analyzed in the study.. The covariate x_i can be the actual size or a vector of categorical variables indicating whether the tumor is in that size category or the excluded baseline category (e.g., medium versus small, large versus small). If tumor-size is not included in the analysis, x_i is the empty set.

6.4.2 Model for multinomial outcome: generalized logistic regression

The distribution of y_i is multinomial with five categories:

$$y_i|\theta_i \sim Multinomial_5(1, \theta_i)$$

where $\theta_i = (\theta_{i0}, \theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4})$ is the vector of probabilities associated with the five possible outcomes for patient *i*. Models for multinomial outcomes are characterized by the way in which the elements of θ_i are mapped to the predictors. One commonly used specification is generalized logistic regression in which each component of θ_i is compared to all other components through a logit link function by supposing:

$$\theta_{ik} = \lambda_{ik}/(\lambda_{i0} + \lambda_{i1} + \lambda_{i2} + \lambda_{i3} + \lambda_{i4})$$

and $\log(\lambda_{ik}) = \gamma_k + \beta x_i$ for $k = 0, \ldots, 4$. A key feature of this specification is that the Bayesian software package JAGS will impute any y_{ik} whose value is missing. To implement this feature any values of y_{ik} that are missing will be represented as NA in the analytic data set. We note that the above specification assumes the proportional odds assumption that the effects of all terms other than γ_k does not depend on *k* holds. This is an assumption whose legitimacy can be tested empirically.

6.4.3 Prior distribution

To complete the specification of the Bayesian model we specify prior distributions for the model parameters. We assume $\gamma_k \sim Normal(0, 10^6)$ and $\beta \sim Normal(0, 10^6)$. These priors impose essentially no information on the analysis due to the extremely large variances of their distributions. If the categorical representation of tumor size is used then a multivariate prior will be assumed for β.

6.4.4 Evaluation of hypothesis test and other summaries of the posterior distribution

We wish to average the 2-year survival probability across the study patients, given by

$$q_{\mathrm{RFA}} = 1 - n^{-1} \sum_{i=1}^{n} \theta_{i4}$$

where based on the above generalized logit model the probability of clinical success at 2-years for patient *i* is given by

$$\theta_{i4} = \frac{\exp(\gamma_4 + \beta x_i)}{\exp(\gamma_0 + \beta x_i) + \exp(\gamma_1 + \beta x_i) + \exp(\gamma_2 + \beta x_i) + \exp(\gamma_3 + \beta x_i) + \exp(\gamma_4 + \beta x_i)}$$

The posterior probability of clinical success, $\Pr(q_{\mathrm{RFA}} < q_{\mathrm{base}} \mid y_1, \ldots, y_n, x_1, \ldots, x_n)$, is then determined by evaluating q_{RFA} for a large number of draws from the posterior distribution of the model parameters and evaluating the Monte-Carlo average

$$\Pr(q_{\mathrm{RFA}} < q_{\mathrm{base}} | \, y_1, \ldots, y_n, x_1, \ldots, x_n) \cong n^{-1} \sum_{j=1}^{n_{sim}} I(q_{RFA}^{j} < q_{base})$$

where q_{RFA}^{j} denotes the value of q_{RFA} computed on the j'th of n_{sim} draws of the model parameters from their joint posterior distribution. The mean, standard deviation, and 95% interval estimates of q_{RFA} will be similarly determined.

Because the chance of clinical success likely varies continuously with tumor size, we can substitute the mean tumor size for the benchmark population in [4] for x_i to essentially control for tumor size. Alternatively, we may substitute x_i with the corresponding small, medium, and large tumor size subgroup mean and in so-doing evaluate the posterior probability with closer control for the tumor size distribution. For the subgroup analyses that compare the performance of RFA-GP on the small, medium and large tumor

subgroups we may estimate $\Pr(q_{\text{RFA}} < q_{\text{base}} \mid \text{size} = s, y_1, \ldots, y_n, x_1, \ldots, x_n)$ separately for each subgroup $s \in \{\text{small}, \text{medium}, \text{large}\}$.

6.4.5 Technical Note: Alternative computational strategy

If there are any technical challenges getting the generalized logistic regression model to run in JAGS (the Bayesian statistical software) with missing values for some of the multinomial outcomes, we will use the continuation-ratio model defined by the logits of the events: recurrence-free versus recurrence at 6 months, recurrence-free versus recurrence at 12 months given recurrence-free to 6 months, recurrence-free versus recurrence at 18 months given recurrence-free to 12 months, and recurrence-free versus recurrence at 24-months given recurrence-free to 18-months. Thus, each successive analysis is conditioned on a reduced set of subjects. It can be shown that the multinomial likelihood function decomposes into separate likelihood functions for each of the above conditional contrasts. Importantly, it is computationally easier to estimate separate logistic regression models as opposed to simultaneously needing to estimate the generalized logit model.

The probabilities estimated by the respective logits of the continuation ratio model are given by:

$$\eta_{i0} = \theta_{i1} + \theta_{i2} + \theta_{i3} + \theta_{i4}$$
$$\eta_{i1} = (\theta_{i2} + \theta_{i3} + \theta_{i4})/(\theta_{i1} + \theta_{i2} + \theta_{i3} + \theta_{i4})$$
$$\eta_{i3} = (\theta_{i3} + \theta_{i4})/(\theta_{i2} + \theta_{i3} + \theta_{i4})$$
$$\eta_{i4} = \theta_{i4}/(\theta_{i3} + \theta_{i4})$$

or equivalently

$$\theta_{i0} = 1 - \eta_{i1}$$
$$\theta_{i1} = \eta_{i1}(1 - \eta_{i2})$$
$$\theta_{i2} = \eta_{i1}\eta_{i2}(1 - \eta_{i3})$$
$$\theta_{i3} = \eta_{i1}\eta_{i2}\eta_{i3}(1 - \eta_{i4})$$
$$\theta_{i4} = \eta_{i1}\eta_{i2}\eta_{i3}\eta_{i4}$$

Because there is a one-to-one relationship between the sets of probabilities, following estimation of the separate logistic regression models we simply need to multiply the necessary probabilities together to recover the probabilities for the multinomial model and thus evaluate the hypothesis test for clinical success.

6.4.6 Power for clinical success

In lieu of performing an exact power calculation under the Bayesian model specified above, we will perform an illustrative calculation based on Fisher's exact test. Allowing a type I error of 0.05, the test of the above one-sided hypotheses for q_{RFA} with a sample-size of 62 lesions will reject the null hypothesis with 80% power if $q_{\text{RFA}} < 0.105$. That is, the minimal detectable difference is a reduction of approximately 0.117 in the likelihood of tumor reoccurrence – which is an effect size in line with conservative hypotheses about the improvement in outcomes that the PGP might introduce.

Subgroup analyses (e.g., based on treated tumor size) will require larger effect sizes to retain the same level of power. However, the study is only intended to be powered for the primary analysis.