Identification of Prognostic Factors for Survival in Neuroendocrine Tumor Patients in the Presence of Multivariate Missingness

Evan Walser-Kuntz and Philip Stallworth

Outline

- Background
  - What is NET?
  - Why it matters
- Our Project
  - Project goals
  - Lymph Node Ratio and Tumor Size
  - Problems with the Data
- Multiple Imputation
  - The Basics
  - The Process
- Statistical Techniques
  - Kaplan-Meier
  - Model Comparison
- Results, Findings, and Recommendation
- Acknowledgements and References

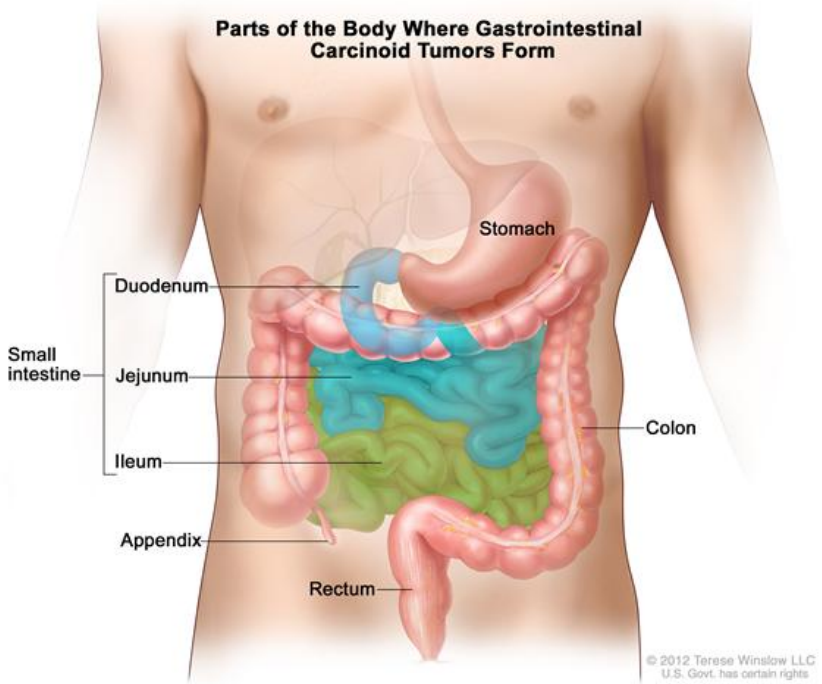
Background

- Neuroendocrine tumors (NET) are neoplasms which arise from cells of the endocrine (hormonal) and nervous systems. (Vinik, et. al 2012)
- NET Carcinoids are a specific type of tumor.
  - Rare, typically small, and have a slow progression
  - Often misdiagnosed because they only become symptomatic after metastasis to the liver and bone
  - Over half occur in the small intestine, but occasionally they can be found in the lungs and the pancreas

Why it Matters

- Although NET has a fairly low prevalence it is still very lethal and the number of cases is enough to warrant concern.
  - 50,000 cases in the United States
  - 1.5 new cases per 100,000(2500 cases per year)
  - NET carcinoids account for 13%-34% of tumors of the small bowel and 17%-46% of all those which are malignant.
  - Overall, there is a moderate to high chance of metastasis for these tumors. Though, the risk relates to both location and size.

Neuroendocrine Tumor Locations



http://www.cancer.gov/images/cdr/live/CDR741567-571.jpg

Project Goals

- Doctors are interested in knowing whether tumor size and Lymph Node Ratio (LNR) are effective prognostic factors for survival in NET patients when added to a list of already established factors.
- Currently, doctors know some prognostic factors for the survival of patients with NETs.
  - *Surgery*: The best prognostic factor to date
  - *Grade*: How much of the cancer has spread to the liver
  - *Stage*: Has the tumor metastasized, spread regionally, or remained local
  - *Location*: Where the tumor originated
  - *Age*: Younger patients are more likely to survive
  - *Marital Status*: Married patients may feel as though they have a strong reason to survive

Lymph Node Ratio and Tumor Size

- Cancer will often first spread to the lymph nodes, so measuring the LNR helps doctors understand how the cancer has progressed.
- To find a patient's LNR, doctors extract lymph nodes during surgery. The number of cancerous lymph nodes divided by the total number extracted yields LNR.
- Doctors believe tumor size may be an important prognostic factor for certain sites.
- Large tumors are those which are bigger than 2 cm. The rate of metastases in these tumors is 95%, compared to 15% in smaller ones (around 1cm). It should follow that size predicts survival.

Our Data

- Our project has the advantage of access to a large database.

- Our analyses used the Surveillance, Epidemiology, and End Results (SEER) dataset published by the National Cancer Institute which pools thousands of cases from areas all over the United States.

Problems with the Data

- Grade, tumor size, and LNR contain high levels of missingness. Surgery and Age have a small degree of missingness.
- High levels of missingness make complete case analysis unfeasible. In our case, we would lose about 75% of our information.

Variable	**Missingness Proportion**
Grade	0.7404
Tumor Size	0.2747
LNR	0.485

What can we do?

- Multiple Imputation (MI), developed by Rubin(1987), to fill in the missing data.
- MI can be summed up by the following process:
    1. Take note of your complete and missing data
    2. Fill in all the missing data from a probability distribution
    3. Repeat steps 1 and 2 *m* times and pool them together to estimate parameter of concern.

$$P(Q|Y_{obs}) = \int P(Q|Y_{mis}, Y_{obs})P(Y_{mis}|Y_{obs})\, dY_{mis}$$

- This process requires the data to be missing at random (MAR) and distinct.

Assumptions of Multiple Imputation

- **Missing at Random (MAR)**: The missing data mechanism does not depend on the missing data. It still may depend on the observed data.
- **Distinct**: The missing data model($P(Y_{mis}|Y_{obs})$) and the analysis model($P(Q|Y_{mis}, Y_{obs})$) are independent.
- **Number of Imputations**: Theoretically, it would be best to perform a very large number of imputations. However, this is neither necessary nor practical because it wastes computational resources.

Relative Efficiency

Given some proportion of missingness, λ, and a finite number of imputations, m, we can compute the relative efficiency, $RE = (1 - \frac{\lambda}{m})^{-1}$, compared to an infinite number of imputations.

λ

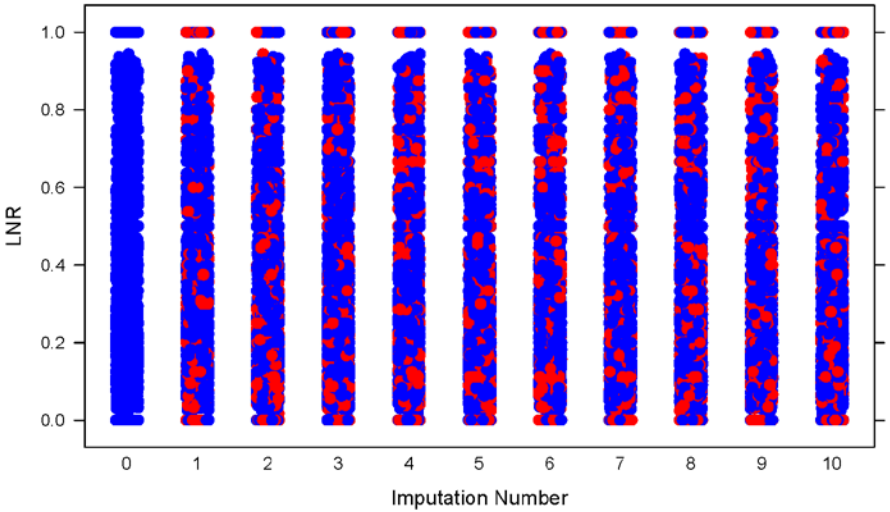
m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346

Imputation Models

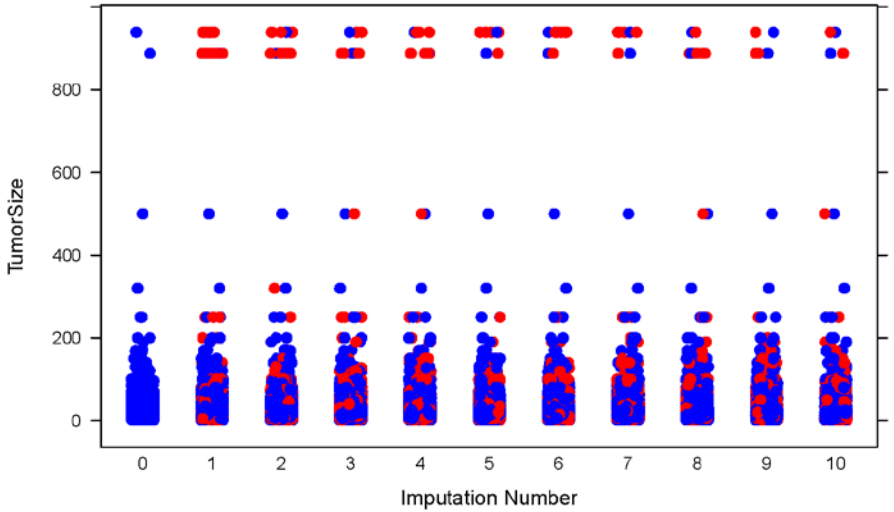
- We used 3 separate techniques:
  - Predictive Mean Matching - Continuous variables
  - Logistic Regression - Dichotomous categorical variables
  - Poly-Logit Regression - Categorical variables with more than two levels
- **Predictive Mean Matching:** Suppose you have k variables $y_1, y_2, \ldots, y_k$ and y_k is both continuous and the only variable with missingness. We regress $y_k \sim \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_{k-1} y_{k-1}$ using estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$ obtained through complete case analysis. Suppose we are interested in finding the ith case for the kth variable, but that observation is missing. We estimate y_{ik} using $\hat{y}_{i,k} = \hat{\beta}_0^* + \hat{\beta}_1^* y_{i,1} + \hat{\beta}_2^* y_{i,2} + \cdots + \hat{\beta}_{k-1}^* y_{i,k-1}$. However, this method may result in odd values. So instead of using this exact number, we find observed data $y_{j,k}$ that are "close" to our estimate. Then, we randomly choose one of those values as our imputed datum for $y_{i,k}$.
- Logistic regression and poly-logit regression perform an analogous task for factored variables.
- Our data
  - Predictive Mean Matching: Lymph Node Ratio, Tumor Size
  - Poly-Logit Regression: Grade, Age
  - Logistic Regression: Surgery Status

Strip-Plot Validation

Strip Plot for LNR



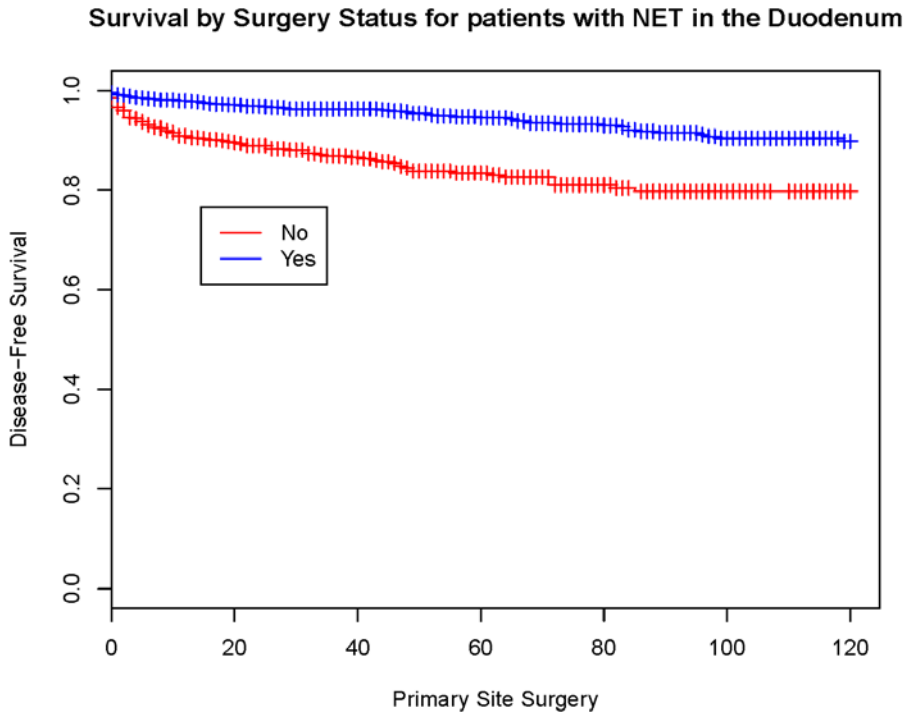
Strip Plot for Tumor Size



Univariate Analyses

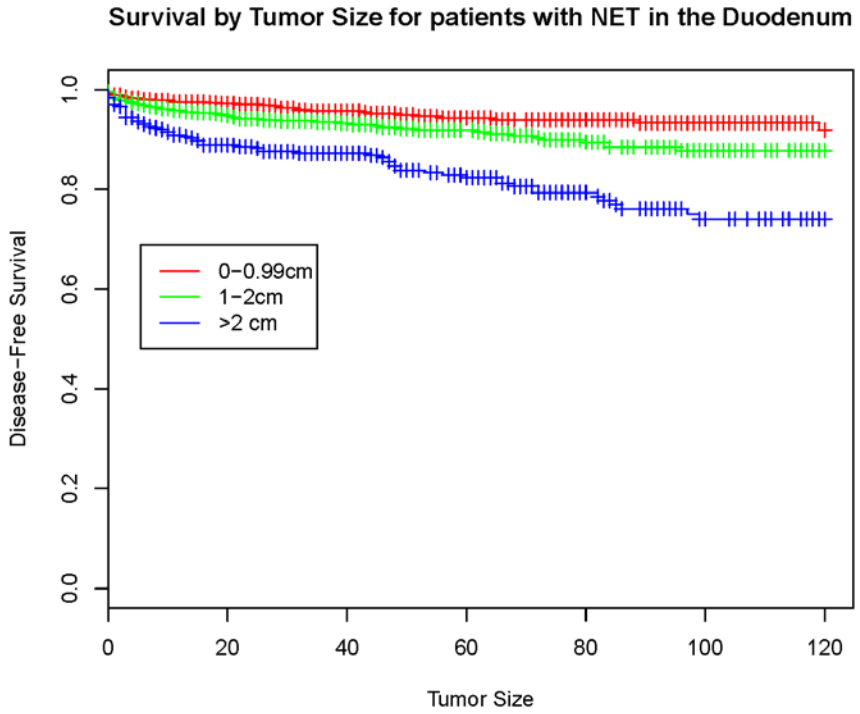
As an exploratory analysis we performed univariate analyses on the imputed data to assess the association between individual factors and survival.

Kaplan Meier: Surgery



Kaplan Meier: Age

Kaplan Meier: Tumor Size



The Cox-Proportional Hazard Model

- Since we are interested in survival and have access to censoring variables we use a Cox-Proportional Hazard model to develop a parsimonious survival model.
- Our process:
  1. Develop a multivariate model with variables currently used to predict the hazard of death for people with NETs.
  2. Add LNR and tumor size to the model
  3. Determine whether their addition provides additional information about the hazard of death by comparing the original model to the model including LNR and tumor size
  4. If it does, we test whether they both contribute or only one does.

Duodenum

Variable		*HR*	*Lower*	*Upper*	*P-Value*
Age	<60				
	60-69	1.75	1.356	2.259	0
	70-79	2.626	2.066	3.337	0
	>80	4.624	3.542	6.036	0
Surgery Status	No Surgery				
	Surgery	0.609	0.476	0.777	0
Stage	Localized				
	Distant	3.106	2.184	4.418	0
	Regional	1.423	0.949	2.133	0.087
	Unstaged	1.007	0.777	1.305	0.959
Marital Status	Not Married				
	Married	0.759	0.636	0.904	0.002
Sex	Male				
	Female	0.764	0.641	0.911	0.003
Tumor Size	≤ 1 cm				
	>1 and ≤ 2 cm	1.053	0.806	1.374	0.701
	>2	1.253	0.911	1.723	0.161
LNR		1.178	0.778	1.781	0.433

Jejunoileal

Variable		*HR*	*Lower*	*Upper*	*P-Value*
Age	<50				
	50-59	1.893	1.457	2.46	0
	60-69	3.24	2.527	4.153	0
	70-79	6.007	4.692	7.69	0
	>80	13.058	9.963	17.114	0
Surgery Status	No Surgery				
	Surgery	0.514	0.407	0.648	0
Stage	Localized				
	Distant	2	1.619	2.472	0
	Regional	0.982	0.799	1.208	0.866
	Unstaged	1.206	0.768	1.892	0.416
Marital Status	Not Married or Widowed				
	Married	0.779	0.674	0.901	0.001
	Widowed	0.764	0.63	0.927	0.006
Tumor Size	≤ 1 cm				
	>1 and ≤ 2 cm	0.969	0.779	1.205	0.775
	> 2 cm	1.006	0.803	1.26	0.96
LNR		1.345	1.066	1.696	0.013

Deviance

We used the change in deviance test to compare the sub-models with and without LNR and tumor size.

Site	P-value for model comparison
Small Bowel, NOS	0.8829
Duodenum	0.3884
Jejunoileal	0.0601
Jejunoileal (No Tumor Size)	0.0149

Our Findings & Recommendations

We found both Lymph Node Ratio and Tumor Size correlated with survival rates in a univariate analysis.

In multivariate analysis, however we found:
In the jejunum and ileum, LNR adds additional information about the hazard of death after considering the known prognostic factors. This is not the case in the duodenum.

We recommend that doctors incorporate LNR into the current prognostic survival model for patients with NETs located in either the jejunum or ileum.

Future Work

- Develop a non-parametric imputation model that does not rely on assumptions
- Learn more about model selection

Acknowledgements

Dr. Gideon Zamba
Dr. T. O'Dorisio
Sarah Bell
National Heart and Lung Association for the grant to participate in ISIB

References

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York.

van Buuren S (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Press, Boca Raton, FL. 342 pages. ISBN 9781439868249

Mamikunian, Gregg, Aaron Vinik, Eugene Woltering, Thomas O'Dorisio, and Vay Liang Go. Neuroendocrine Tumors: A Comprehensive Guide to Diagnosis and Management. Inglewood, CA.: Inter Science Institute, 2009. Print.