Multi-Feature Method for Ascertaining NET Cases in a Clinical Data Warehouse

Marie Synnestvedt PhD[1], Bonnie Bennett RN[1], Hilary Faust MD[1], Mark Weiner MD[1], David C. Metz MD[1]

1. The Perelman School of Medicine at University of Pennsylvania, Philadelphia PA 19104

Abstract

Multi-Feature Method for Ascertaining NET Cases in a Clinical Data Warehouse

Marie Synnestvedt PhD[1], **Bonnie Bennett RN[1]**, Hilary Faust MD[1], Mark Weiner MD[1], David C. Metz MD[1]

The Perelman School of Medicine at University of Pennsylvania, Philadelphia PA 19104

Background: A multidisciplinary program should be the standard of care for neuroendocrine tumors (NETs) but care is often fragmented across multiple specialties. We developed a method to ascertain how many NET patients existed in our health care system in order to establish a registry for future clinical care and research.

Methods: A knowledge discovery approach was undertaken to identify a trial set of criteria by examining terminologies within electronic NET resources such as NANETS and SNOMED CT terminology and performing statistical and text analysis of a sample of 364 known NET cases (128 GI NETs and 236 pheochromocytoma/paragangliomas). The criteria selected included laboratory (any abnormal 5HIAA, chromogranin A, epinephrines, metanephrines, norepinephrines, pancreatic polypeptides, serotonin, or VMA), pathology (positive keywords in diagnosis), radiology (non-negative MIBG or Octreoscan), medications for pheochromocytoma, or ICD9 codes (for genetic diagnoses such as MEN). These criteria yielded 4789 potential NET cases going back to 1999. A 10% random sample of 274 cases was selected from 2652 unclassified cases with activity since 2008 for review and classification of true NET status.

Results: The search method retrieved 89% of known NET cases. Missed cases were primarily older cases that predated searchable electronic data or that were diagnosed using historical terminology such as glomus tumor. The true case rate in the random sample was 132/274 (48%) 56 GI NETs, 14 pheochromocytomas or paragangliomas, 27 lung carcinoids and 35 other NETs (142 were not NETs). Decision tree classification modeling was applied to the labeled random sample. The resulting model predicts NE vs non-NET patient records with sensitivity of 77%, specificity of 86%, positive predictive value of 83% and negative predictive value of 80%.

Conclusion: This methodology can be improved with further modeling, and in its current form can help institutions identify a larger cohort of previously unidentified patients with NETs who are suitable for inclusion in a registry for follow-up, management, and improved research capabilities.

Aims

- Identify NET patients in a large academic health system to create a registry for future study and quality improvement
- Identify search terms most sensitive and specific for NET patients
- Develop a model that accurately detects NET patients in the health system
- Assess accuracy of model to validate it for future identification of NET patients

Methods

- Examined NANETs, SNOMED and patient records to assemble list of search terms (Table 1)
- Applied search terms to health system data as preliminary screen for potential NET patients
- Manually reviewed a 10% random sample of recent cases to classify true NET status
- Use decision tree analysis to develop predictive model using the random sample of confirmed NET and non-NET cases

Results

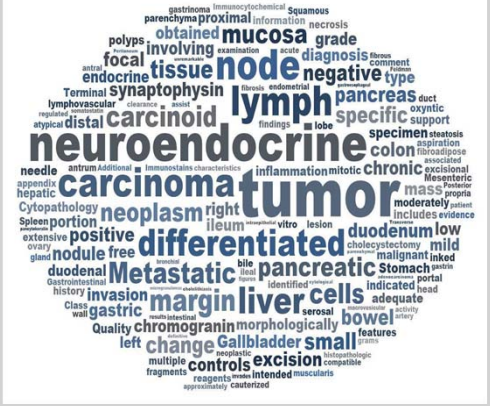
Results: Search Terms

Table 1: Search Criteria for Retrieval of Data

Diagnosis Codes for Genetic NETs	237.71 Neurofibromatosis, Type 1 (Von Recklinghausen's Di 258.01 Multiple Endocrine Neoplasia (Men) Type I 258.02 Multiple Endocrine Neoplasia (Men) Type IIa 258.03 Multiple Endocrine Neoplasia (Men) Type IIb 759.6 Other Congenital Hamartoses, (VhI (Von Hippel-Lindau Syndrome) V84.81 Genetic Susceptibility To Multiple Endocrine Neoplasia
Imaging	MIBG sc an Octreoscan
Laboratory	5HIAA Chromog Epinep Metanep Norepinep Pancpoly Serotonin Umetanep UVMA
Medication	Metyrosine Phenoxybenzamine hcl
Pathology Diagnosis Terms	Apudoma Carcinoid Ganglioneuro Gastrinoma Glomus Glucagonoma Insulinoma Medullary carcinoma Neuroendocrine Pancreatic endocrine Paraganglioma Phenochromocytoma Pheochromocytoma Somatostatinoma Synaptophysin

A trial set of search criteria selected using clinical resources such as NANETS and SNOMED CT [1] included ICD9 codes, common laboratory evaluations, imaging studies, medications and pathology studies frequently and uniquely associated with neuroendocrine tumors. These criteria were applied to the data warehouse and identified 4799 potential NET cases.

Figure 1: Word Cloud of Pathology Search Terms



Word cloud of terms derived from text of pathology diagnosis for known NET patients, with word size correlating with relative frequency of terms.

Results: Demographics

Table 2: NET Patients

Sex	#	%
F	78	59.1%
M	54	40.9%
Total	132	100.0%

Race	#	%
AAB	25	18.9%
ASIAN	5	3.8%
OTHER	12	9.0%
WHITE	80	60.6%
Unknown	10	7.5%
Total	132	100.0%

MinOfAge	MaxOfAge	AvgOfAge
13	88.3	55.96

Demographics of NET patients

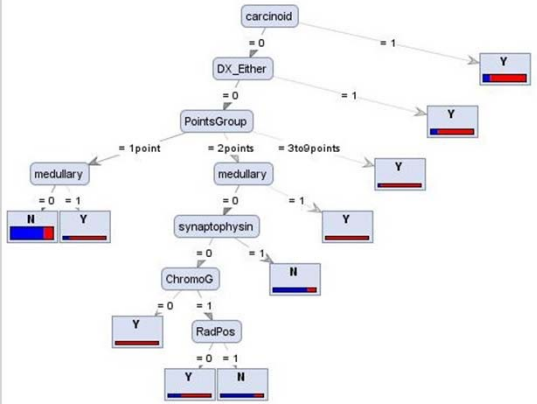
Table 3: NETs By Type

SubSite	Description	#	%
1A	Pheochromocytoma / paraganglioma adrenal / extra-adrenal	6	4.5%
1B	Pheochromocytoma / paraganglioma other	8	6.1%
2A	GI NETS PNET	19	14.4%
2B1	GI NETS Alimentary tract - foregut	7	5.3%
2B2	GI NETS Alimentary tract – mid gut	18	13.6%
2B3	GI NETS Alimentary tract – hind gut	12	9.1%
3	Lungs and Thymus including mediastinum	27	20.5%
4A	Other tumor Medullary Thyroid Cancer	9	6.8%
4B	Other tumors Pituitary	5	3.8%
4C	Other tumors Others (testes, heart,..)	1	0.8%
5	Genetic disorders with a predisposition for developing neuroendocrine tumors No tumors yet MEN 1 MEN 2 VHL NF – 1	20	15.2%
Total		132	100.0%

Types of neuroendocrine tumors ascertained using multi-feature method

Results: Model Derivation and Performance

Figure 2: Training-Sample Derived Model



Decision tree analysis model derived from labeled random sample. Each node of the tree labeled "Y" represents a group of cases predicted as being a NET case, and the height of the bar represents the relative number of cases in the tree. Sections of the bar that are colored red indicate the proportion of true NET cases within the node. The "PointsGroup" variable refers to the total number of positive criteria found for a case.

Tables 3 and 4: Performance of Model

		Actual		
		Non-NET	NET	Totals
Predicted	NET	20	101	121
	Non-NET	122	31	153
	Totals	142	132	274

Confusion matrix demonstrating the model's predicted NET and non-NET totals vs the actual proportions of NET and non-NET cases in the sample. There were 20 false positives and 31 false negatives in a sample of 132 true positive and 142 true negative NETs.

	Estimated Value	*95% Confidence Interval*	
		Lower Limit	*Upper Limit*
Sensitivity	.77	.68	.83
Specificity	.86	.79	.91
Positive Predictive Value	.83	.75	.89
Negative Predictive Value	.80	.73	.86

Table demonstrating the accuracy of the model developed on a sample of confirmed NETs and non-NETs.

Summary

- An initial screen of health system data using 34 search terms **identified 4789 patients**
- A **decision tree analysis model** for identifying NET patients was developed using a labeled sample of NET and non-NET cases
- The model yielded **sensitivity of 77%, specificity of 86%, positive predictive value of 83%** and **negative predictive value of 80%**

Conclusions

- **Models can be developed to identify current NET patients in a large health system with relative accuracy and efficiency over current methods of identifying NET patients**
- **These models can be improved over time by adding more search terms and by validated them on larger samples of potential NET patients**
- **Accurate identification will facilitate future study of patients with neuroendocrine tumors**

References

1. International Health Terminology Standards Development Organisation - SNOMED CT. http://www.ihtsdo.org/snomed-ct/