Computable Phenotypes for Identifying Patients with Lung and Gastroenteropancreatic Neuroendocrine Tumors

Table of Contents

Overview

The Neuroendocrine Tumors – Patient Reported Outcomes (NET-PRO) study uses a patient-centered, pragmatic approach to evaluate treatment effects in lung and gastroenteropancreatic (GEP) neuroendocrine tumors (NETs). This observational study leverages PCORnet, a network encompassing 14 healthcare systems across the US.

The NET-PRO teams at each site will contact potentially eligible patients by electronic/ground mail or will approach them in person at the clinic. Understandably, each site may prefer a particular form of contact that will optimize their enrollment.

In order to accommodate these varied approaches, we have developed a multi-pronged approach that leverages existing data sources to identify potential enrollees. They include:

- Phenotypes that use diagnosis codes in electronic health record (EHR) data in clinical data research warehouses or most recent PCORnet Common Data Model (CDM) refresh.
    - A phenotype is provided for using these data for **low-touch** recruitment (e.g., by e-mail). This algorithm is designed to maximize positive predictive value.
    - A phenotype is provided as a **first pass** for identifying potential recruits. Eligibility for patients identified with this tool should be confirmed before they are approached. This algorithm is designed to maximize sensitivity.

- Phenotypes that use data from institutional **tumor registries**
    - These data have a typical lag time of one year (depending on the institution), but an excellent phenotype is provided for using this data source as part of a low-touch recruitment strategy

This document describes the advantages and disadvantages of each data source and provides technical specifications for computable phenotypes that are appropriate for different recruitment strategies. For example, study sites that plan to recruit from their clinics or conduct chart reviews before recruiting may find it most effective to use the first pass phenotype to identify potential participants and verify eligibility through chart review; staff could screen out the false positives that would be flagged with such a phenotype. Sites that plan to recruit via e-mail or ground mail *without prior chart review to confirm eligibility*, however, should use the low touch or tumor registry phenotypes to maximize the possibility that contacted patients actually have a neuroendocrine/carcinoid tumor (NET) diagnosis. (There is generally a trade-off between PPV and sensitivity; higher PPV is associated with lower sensitivity and higher sensitivity is associated with lower PPV.)

The validity of each computable phenotype is also discussed. Where available, results from validation studies performed at the University of Iowa Coordinating Center (UICC) are provided. Each site is encouraged to conduct its own validation studies.

It is useful to remember that prospective enrollees will confirm their eligibility by responding to eligibility questions early in the enrollment process. This will serve as a final check in case the phenotype mistakenly identifies someone as having a qualifying NET. The recruiting materials have been worded carefully to minimize the possibility that a potential enrollee is alarmed to be contacted for a study focused on NETs.

Phenotypes that use diagnosis codes in electronic health record (EHR) data (research warehouse or CDM)

Many research institutions maintain medical record and billing data in a clinical research data warehouse (CRDW) that contains structured information on diagnoses and procedures. The PCORnet Common Data Model (CDM) is often populated by this source data. Computable phenotypes can be applied to the CDM or the underlying CRDW (which may be updated more frequently than the CDM).

An advantage of CRDW data is that patient data are structured by clinical or billing staff in real-world settings using the International Classification of Diseases (ICD) system (9^{th} and 10^{th} editions). While data are generally high quality, errors can occur. Also, diagnoses can change or be refined over time, and these changes are not always clearly documented. Some ICD codes are ambiguous with regard to tumor site (see the ICD codes labeled AMBIGUOUS in the Appendix). It is also important to note that the date a cancer ICD diagnosis code first appears on an encounter cannot necessarily be considered to be the date of cancer diagnosis; it generally lags behind the true diagnosis date. That lag in time can be substantial, especially if a patient was originally diagnosed or treated at another institution.

Phenotype for **low-touch** recruitment

- This phenotype was designed to maximize positive predictive value so it would be appropriate for recruitment via electronic or ground mail without prior confirmation of eligibility through chart review or clinic visit.
- The phenotype identifies patients with at least one ICD code in their electronic medical record (EMR)/billing records that specifies a GEP or lung NET. To mitigate the effect of incidental coding errors, at least two NET codes (which may or may not specify the site of the NET; see code descriptions in the Appendix) must be present, and the patient's first and last NET code must be more than 30 days apart.
- The first NET code for a patient is presently required to be dated on or after 01JAN2018. This increases the likelihood that the patient's NET diagnosis will be in the required 2018-2023 study window.
- This phenotype should be applied for all patients who are not known to have died.
- The lung NET High PPV phenotype had a PPV of 92% and the GEP NET High PPV phenotype had a PPV of 90-98% when compared with chart review at The University of Iowa. The estimated sensitivity was 59.0% and 45.2%, respectively (see Technical Details Box 1 for further information)

Technical Details Box 1. Performance of Phenotypes that use Diagnosis Codes in EHR Data (research warehouses or CDM)

1. Low Touch Phenotype

Estimating the Positive Predictive Value of the Low Touch Phenotype

The University of Iowa (UI) team conducted a chart review to validate a 15% random sample of cases identified with this phenotype over a five-year period. This was done separately for lung (N = 13) and GEP NETs (N = 51).

- Of the 13 cases identified by the phenotype as having lung NETs, 12 had diagnoses that were unequivocally confirmed on chart review (i.e., the chart described pathological confirmation of a lung NET). This corresponds to a PPV of 92%. The single case that was not confirmed had a pancreatic NET.
- Of the 51 cases identified by the phenotype as GEP NETs, 46 had diagnoses that were unequivocally confirmed on chart review. This corresponds to a PPV of 90%. There were two other cases that had pathological evidence of a NET with an unknown primary location. An additional two cases had a suspected NET with no pathological confirmation on the chart. If these four cases are considered "hits", PPV increases to 98%. For the one remaining case, the chart contained contradictory and ambiguous information about the patient's tumor.

Estimating the Sensitivity of the Low Touch phenotype

The UI team used the University of Iowa Hospitals and Clinics (UIHC) institutional tumor registry to identify cohorts of patients with confirmed GEP and lung NETs (N = 166 and N = 39, respectively). These cases were diagnosed over a two-year period. The phenotype was applied to these cohorts to obtain a measure of sensitivity.

- Of the 39 cases with lung NETs, 23 cases were identified with the phenotype. This corresponds to sensitivity of 59.0%.
- Of the 166 cases with GEP NETs, 75 cases were identified with the phenotype. This corresponds to sensitivity of 45.2%.

2. First Pass Phenotype

Estimating the Sensitivity of the First Pass Phenotype

As above, the UI team used the UIHC institutional tumor registry to identify separate cohorts of patients with confirmed GEP and lung NETs (N = 166 and N = 39, respectively). The first pass phenotype was applied against these cohorts to obtain a measure of sensitivity.

- Of the 39 cases with lung NETs, 33 cases were identified with the phenotype. This corresponds to sensitivity of 84.6%.
- Of the 166 cases with GEP NETs, 148 cases were identified with the phenotype. This corresponds to sensitivity of 89.2%.

Implementation for low touch phenotype

Defintions

```
%LET GEPCODES10 = 'C7A.01','C7A.010','C7A.011','C7A.012','C7A.019','C7A.020','C7A.021','C7A.022','C7A.023',
'C7A.024','C7A.025','C7A.026','C7A.029','C7A.092','C7A.094','C7A.095','C7A.096','C25.4';
%LET GEPCODES09 = '209.00','209.01','209.02','209.03','209.10','209.11','209.12','209.13','209.14',
'209.15','209.16','209.17','209.23','209.25','209.26','209.27','157.4';

%LET LUNGCODES10 = 'C7A.090';
%LET LUNGCODES09 = '209.21';

%LET ambiguousNET10 = 'C7A.00','C7A.098','C7A.1','C7A.8';
%LET ambiguousNET09 = '209.20','209.29','209.30';
```

Criterion	Logic/Notes	Defined
1. At least one ICD code that specifies a GEP or lung NET	Basis: CDM DIAGNOSIS table or source data equivalent	At least one record for each patient: Lung NETs (DX_TYPE = '09' AND (DX IN (&LUNGCODES09))) OR (DX_TYPE = '10' AND (DX IN (&LUNGCODES10))); GEP NETs (DX_TYPE = '09' AND (DX IN (&GEPCODES09)))

		OR (DX_TYPE = '10' AND (DX IN (&GEPCODES10)));
2. Patient's first NET code dated on or after 01JAN2018	Basis: CDM DIAGNOSIS table or source data equivalent	Define first_net_dt: <u>Lung NETs</u> [set of lung/ambiguous site NET records] → (DX_TYPE = '09' AND (DX IN (&LUNGCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&LUNGCODES10,&ambiguousNET10))); <u>GEP NETs</u> [set of GEP/ambiguous site NET records] → (DX_TYPE = '09' AND (DX IN (&GEPCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&GEPCODES10,&ambiguousNET10))); first_net_dt = of the set of lung/GEP NET records for each patient, min(admit_date) ***

		first_net_dt >= '01JAN2018'd
3. Days between patient's first and last NET code is greater than 30 days	Basis: CDM DIAGNOSIS table or source data equivalent	Define first_net_dt and last_net_dt: <u>Lung NETs</u> [set of lung/ambiguous site NET records] → (DX_TYPE = '09' AND (DX IN (&LUNGCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&LUNGCODES10,&ambiguousNET10)));

		GEP NETs [set of GEP/ambiguous site NET records] → (DX_TYPE = '09' AND (DX IN (&GEPCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&GEPCODES10,&ambiguousNET10))); first_net_dt = of the set of lung/GEP NET records for each patient, min(admit_date) last_net_dt = of the set of lung/GEP NET records for each patient, max(admit_date) *** if intck('day',first_net_dt,last_net_dt) > 30

First Pass Phenotype for use when resources can confirm eligibility first, e.g. chart review or in clinic recruitment

- This phenotype was designed to maximize sensitivity so it would be appropriate for clinic recruitment or chart confirmation prior to e-mail or ground mail recruitment.
- The phenotype relaxes the criteria described above and identifies patients with at least one ICD code in their EMR/billing records that specifies a NET (see codes in the Appendix).
- Relaxing the number of codes required resulted in a sensitivity of 84.6% (lung) and 89.2% (GEP)(See Technical Details Box 1).

Implementation for first pass phenotype

Criterion	Logic/Notes	Defined
1. At least one ICD code that specifies a NET	Basis: CDM DIAGNOSIS table or source data equivalent	At least one record for each patient: <u>Lung NETs</u> (DX_TYPE = '09' AND (DX IN (&LUNGCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&LUNGCODES10,&ambiguousNET10))); <u>GEP NETs</u> (DX_TYPE = '09' AND (DX IN (&GEPCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&GEPCODES10,&ambiguousNET10)));
2. Patient's first NET code dated on or after 01JAN2018		<u>Lung NETs</u> [set of lung/ambiguous site NET records] → (DX_TYPE = '09' AND (DX IN (&LUNGCODES09,&ambiguousNET09))) OR

		(DX_TYPE = '10' AND (DX IN (&LUNGCODES10,&ambiguousNET10))); <u>GEP NETs</u> [set of GEP/ambiguous site NET records] → (DX_TYPE = '09' AND (DX IN (&GEPCODES09,&ambiguousNET09))) OR (DX_TYPE = '10' AND (DX IN (&GEPCODES10,&ambiguousNET10))); first_net_dt = of the set of lung/GEP NET records for each patient, min(admit_date) *** first_net_dt >= '01JAN2018'd

Adding Data from Institutional Tumor Registries to Expand Low-Touch Recruitment

All hospitals at NET-PRO participating sites are accredited by the American College of Surgeons' Commission on Cancer (CoC). To maintain this accreditation, they are required to maintain a tumor registry. These registries identify tumors with malignant behavior and abstract their clinical

characteristics, patient demographics, and treatment. This work is performed in compliance with the standards developed by the North American Association of Central Cancer Registries (NAACCR).

Tumor registries document incident, primary cancers. Cases are documented by trained abstractors who gather information directly from patient medical records. Registry records of NETs can be regarded with high confidence. Patients identified through this source could be approached for recruitment through e-mail or similar methods without necessitating prior chart review or clinic visit to confirm eligibility. Documentation is extensive and complies with NAACCR standards. Cancer site and histology are coded according to the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3).

When present in tumor registries, the diagnosis can be considered confirmed. However, there is usually a lag between diagnosis and case abstraction; cases are typically abstracted not less than 4-6 months after diagnosis, and it can take more than a year before a registry abstracts all eligible cases for a given diagnostic year. Another limitation is that tumor registries may prioritize abstracting those cancers that are diagnosed or treated at their respective institutions. Patients who are seen for tests only, a "second opinion", or other consultation may not be abstracted, or their abstraction can be delayed. So, although documentation of a case can be regarded with high confidence, there are likely other eligible cases whose records are not abstracted. At the University of Iowa Hospitals and Clinics (UIHC), we have found that about two-thirds of eligible NET cases are found in tumor registry records. We have also found that about 45-59% of patients detected with the low touch phenotype (described in the previous section) are found in the tumor registry records. Hence, tumor registry data can be potentially useful to increase the number of patients for low touch recruitment, and this phenotype could be used in conjunction with the other phenotypes to increase eligible cases

Phenotype for using tumor registry data as part of a low touch recruitment strategy

- This computable phenotype selects those patients with primary lung or GEP NETs diagnosed between 2018 and 2023 who are 18 years of age and older at diagnosis.
- Cases have been documented by trained abstractors who gather information directly from patient medical records. The information is considered to be of high quality.
- The unique patients identified by the tumor registry algorithm can be directly recruited without need for chart confirmation or clinic visit.

Implementation

Criterion	Logic/Notes	Defined
1. Lung/GEP tumor site	Basis: CDM TUMOR table or	Lung tumor site

	source data equivalent (NAACCR #400 - Primary Site) Sites that use the PCORnet tumor table must establish a link to a table with the identifying data that is necessary for recruitment; the tumor table itself is deidentified. Tumor registries generally do not store the decimal in the ICD-O-3 site codes, but this may vary between hospitals.	PRIMARY_SITE_N400 in ('C340','C341','C342','C343','C348','C349'); <u>GEP tumor site</u> PRIMARY_SITE_N400 in ('C160','C161','C162','C163','C164','C165','C166','C168','C169','C170','C171','C172','C173','C178', 'C179','C180','C181','C182','C183','C184','C185','C186','C187','C188','C189','C199','C209','C250','C251', 'C252','C253','C254','C257','C258','C259','C260','C268','C269');
2. NET histology	Basis: CDM TUMOR table or	HISTOLOGIC_TYPE_ICD_O3_N522 in ('8150','8151','8152','8153','8155','8156','8157','8240','8241','8242','8246','8249');

	source data equivalent (NAACCR #522) Some sites may store histology in the first four characters of the morphology variable (NAACCR #521 Morph--Type&Behav ICD-O-3). This variable is not included in Version 1.2 of the PCORnet tumor table.	
3. Diagnosed between 01/01/2018-12/31/2023	Basis: CDM TUMOR table or source data equivalent (NAACCR #390)	substr(DATE_OF_DIAGNOSIS_N390,1,4) in (‘2018’,'2019', '2020', '2021', '2022', '2023');
4. Aged ≥18 years at diagnosis	Basis: CDM TUMOR table or source data	input(AGE_AT_DIAGNOSIS_N230,3.) >= 18;

	equivalent (NAACCR #230)	

Other Recruitment Methods

Direct recruitment of eligible patients attending clinic will require a liaison with the practice clinics to assess patient appointments and coordinate with the clinical team to introduce the study in-person and supply a study packet/e-mail invitation. Provision of specific guidance on this is difficult, as the process will differ according to the logistics of each site.

Appendix

ICD-O-3 Site Codes

SITE	CODE	DESCRIPTION
GEP	C16.0	Cardia, NOS
GEP	C16.1	Fundus of stomach
GEP	C16.2	Body of stomach
GEP	C16.3	Gastric antrum
GEP	C16.4	Pylorus
GEP	C16.5	Lesser curvature of stomach NOS
GEP	C16.6	Greater curvature of stomach NOS
GEP	C16.8	Overlapping lesion of stomach
GEP	C16.9	Stomach, NOS
GEP	C17.0	Duodenum
GEP	C17.1	Jejunum
GEP	C17.2	Ileum
GEP	C17.3	Meckels diverticulum
GEP	C17.8	Overlapping lesion of small intestine
GEP	C17.9	Small intestine, NOS
GEP	C18.0	Cecum
GEP	C18.1	Appendix
GEP	C18.2	Ascending colon

GEP	C18.3	Hepatic flexure of colon
GEP	C18.4	Transverse colon
GEP	C18.5	Splenic flexure of colon
GEP	C18.6	Descending colon
GEP	C18.7	Sigmoid colon
GEP	C18.8	Overlapping lesion of colon
GEP	C18.9	Colon, NOS
GEP	C19.9	Rectosigmoid junction
GEP	C20.9	Rectum, NOS
GEP	C25.0	Head of pancreas
GEP	C25.1	Body of pancreas
GEP	C25.2	Tail of pancreas
GEP	C25.3	Pancreatic duct
GEP	C25.4	Islets of Langerhans
GEP	C25.7	Other specified parts of pancreas
GEP	C25.8	Overlapping lesion of pancreas
GEP	C25.9	Pancreas, NOS
GEP	C26.0	Intestinal tract, NOS
GEP	C26.8	Overlapping lesion of digestive system
GEP	C26.9	Gastrointestinal tract, NOS
LUNG	C34.0	Main bronchus
LUNG	C34.1	Upper lobe, lung
LUNG	C34.2	Middle lobe, lung
LUNG	C34.3	Lower lobe, lung
LUNG	C34.8	Overlapping lesion of lung
LUNG	C34.9	Lung, NOS

ICD-O-3 Histology Codes

CODE	DESCRIPTION

8150	Pancreatic endocrine tumor, malignant
8151	Insulinoma, malignant
8152	Glucagonoma, malignant
8153	Gastrinoma, malignant
8155	Vipoma, malignant
8156	Somatostatinoma, malignant
8157	Enteroglucagonoma, malignant
8240	Carcinoid tumor, NOS
8241	Enterochromaffin cell carcinoid
8242	Enterochromaffin-like cell tumor, malignant
8246	Neuroendocrine carcinoma, NOS
8249	Atypical carcinoid tumor

ICD-9 and -10 Codes

TYPE	SITE	CODE	DESCRIPTION
ICD-9	GEP	209.00	Malignant carcinoid tumor of the small intestine, unspecified portion
ICD-9	GEP	209.01	Malignant carcinoid tumor of the duodenum
ICD-9	GEP	209.02	Malignant carcinoid tumor of the jejunum
ICD-9	GEP	209.03	Malignant carcinoid tumor of the ileum
ICD-9	GEP	209.10	Malignant carcinoid tumor of the large intestine, unspecified portion
ICD-9	GEP	209.11	Malignant carcinoid tumor of the appendix
ICD-9	GEP	209.12	Malignant carcinoid tumor of the cecum
ICD-9	GEP	209.13	Malignant carcinoid tumor of the ascending colon
ICD-9	GEP	209.14	Malignant carcinoid tumor of the transverse colon
ICD-9	GEP	209.15	Malignant carcinoid tumor of the descending colon
ICD-9	GEP	209.16	Malignant carcinoid tumor of the sigmoid colon
ICD-9	GEP	209.17	Malignant carcinoid tumor of the rectum
ICD-9	GEP	209.23	Malignant carcinoid tumor of the stomach

ICD-9	GEP	209.25	Malignant carcinoid tumor of foregut, not otherwise specified
ICD-9	GEP	209.26	Malignant carcinoid tumor of midgut, not otherwise specified
ICD-9	GEP	209.27	Malignant carcinoid tumor of hindgut, not otherwise specified
ICD-9	GEP	157.4	Malignant neoplasm of islets of langerhans
ICD-10	GEP	C7A.01	Malignant carcinoid tumors of the small intestine
ICD-10	GEP	C7A.010	Malignant carcinoid tumor of the duodenum
ICD-10	GEP	C7A.011	Malignant carcinoid tumor of the jejunum
ICD-10	GEP	C7A.012	Malignant carcinoid tumor of the ileum
ICD-10	GEP	C7A.019	Malignant carcinoid tumor of the small intestine, unspecified portion
ICD-10	GEP	C7A.020	Malignant carcinoid tumor of the appendix
ICD-10	GEP	C7A.021	Malignant carcinoid tumor of the cecum
ICD-10	GEP	C7A.022	Malignant carcinoid tumor of the ascending colon
ICD-10	GEP	C7A.023	Malignant carcinoid tumor of the transverse colon
ICD-10	GEP	C7A.024	Malignant carcinoid tumor of the descending colon
ICD-10	GEP	C7A.025	Malignant carcinoid tumor of the sigmoid colon
ICD-10	GEP	C7A.026	Malignant carcinoid tumor of the rectum
ICD-10	GEP	C7A.029	Malignant carcinoid tumor of the large intestine, unspecified portion
ICD-10	GEP	C7A.092	Malignant carcinoid tumor of the stomach
ICD-10	GEP	C7A.094	Malignant carcinoid tumor of the foregut, unspecified
ICD-10	GEP	C7A.095	Malignant carcinoid tumor of the mid-gut, unspecified
ICD-10	GEP	C7A.096	Malignant carcinoid tumor of the hindgut, unspecified
ICD-10	GEP	C25.4	Malignant neoplasm of endocrine pancreas
ICD-9	LUNG	209.21	Malignant carcinoid tumor of the bronchus and lung
ICD-10	LUNG	C7A.090	Malignant carcinoid tumor of the bronchus and lung
ICD-9	AMBIGUOUS	209.20	Malignant carcinoid tumor of unknown primary site
ICD-9	AMBIGUOUS	209.29	Malignant carcinoid tumor of other sites
ICD-9	AMBIGUOUS	209.30	Malignant poorly differentiated neuroendocrine carcinoma, any site
ICD-10	AMBIGUOUS	C7A.00	Malignant carcinoid tumor of unspecified site

ICD-10	AMBIGUOUS	C7A.098	Malignant carcinoid tumors of other sites
ICD-10	AMBIGUOUS	C7A.1	Malignant poorly differentiated neuroendocrine tumors
ICD-10	AMBIGUOUS	C7A.8	Other malignant neuroendocrine tumors

Version History

Version 2: The original study criteria included those patients who were diagnosed between 2019 and 2023. In August of 2023, this was expanded to include those diagnosed in 2018 as well. This document was updated to reflect this change; all 2019 references were changed to 2018 where appropriate.