# Annotation Guideline

**Guideline for annotating clinical notes for patients with thyroid cancer to extract cancer staging, risk category and relevant information using BRAT**

This annotation guideline provides annotation framework for extracting the cancer staging and risk category directly from clinical notes for patients with thyroid cancer, and relevant elements to refine the staging and risk categories using BRAT. The sources of clinical notes include 1) The Cancer Genome Atlas - Thyroid Cancer (TCGA-THCA) programme; and 2) pseudo clinical notes created by two endocrine surgeons from a leading tertiary endocrine surgery centre in Hong Kong. The types of pseudo clinical notes include operation records and histopathology reports.

**Clinical information to be extracted**

The disease information to be extracted and annotated are listed below.

Summary of the items for extraction

|  |  |
| --- | --- |
| **Category** | **Items** |
| Cancer staging / risk category | - AJCC TNM Cancer staging  - American Thyroid Association (ATA) risk category |
| Factors to refine staging and risk | - Operation date  - Age at diagnosis  - Lymph node involvement  - Distant metastasis involvement  - Total number of lymph node resected  - Total number of lymph node involved  - Margin involved  - Histopathology and its subtype  - Tumor size  - Size of largest involved lymph node  - Vascular invasion (Angionvasion, blood vessel invasion, blood vascular invasion, lymph-vascular invasion, lymphovascular invasion, and angiolymphatic invasion)  - Number of foci of vascular invasion for follicular thyroid carcinoma  - Capsular invasion  - Gross extrathyroidal extension  - Microscopic extrathyroidal extension  - Mutation and the type  - Thyroidectomy (type, completeness of resection from the perspective of surgeon and pathologist) |

**Methods of annotations**

At least two postdoctoral fellows and/or medical students will independently perform annotations. The performance and consistency of the annotations will be checked by observed and expected agreement and Cohen’s kappa statistic. Any disagreement between the annotators will be resolved by discussion among annotators and expert annotator (i.e. clinicians with specialty of endocrine surgery or clinical oncology).

The BRAT (<https://brat.nlplab.org/index.html>) will be used for the annotation task. The output style of the annotation results is in ANN format which could be implemented for the development of named entity framework. There are three types of annotation modalities will be used for annotation, which are 1) entities of specific information, 2) attributes of entities and 3) relations between entities. For the relation, it refers to descriptive (simple relations between entities) relation. Table 2 summarizes the annotation modalities for items to be extracted using the natural language processing models.

The following general rules will be applied for the annotation in all the tasks:

1. No punctuation and space should be annotated.
2. Values presented in Arabic, Roman or English format should be annotated.
3. If there is more than one information for the same concept which are mutually exclusive to each other within the same note (e.g. two histological types of thyroid cancer diagnosed at the same time), this information should be annotated and labelled separately.
4. If the targeted words / phrases are mentioned multiple times but contradict with each other within the same note, only the key entity with the most significant and latest information or with the highest extent (i.e. most severe) should be annotated. If the information is contradicted between gross examination and microscopic examination, the one listed in gross examination should be annotated.
5. For the entities that have not been mentioned in the notes, it will be assumed that the information is missing.
6. Keywords or phrases implied the concept within the same sentence should be annotated. Otherwise, it will be assumed that the information is in different concept and should not be annotated.
7. Any disagreement between the annotators will be resolved by discussion among annotators and expert annotator (i.e. endocrine surgeons or oncologists).

Annotation modalities for extraction of cancer staging, risk categories and relevant information

| **Modality** | **Category** | **Name** | **Description** |
| --- | --- | --- | --- |
| Entity | Temporal | Date | Operation date |
| Age | Age at diagnosis‡ |
| Disease information | AJCC8 | The 8th edition of AJCC cancer staging |
| Tstage | Tumor stage‡ |
| Nstage | Lymph node stage†‡ |
| Mstage | Metastasis stage†‡ |
| LN | Metastatic lymph node involvement and invading location†‡ |
| Meta | Distant metastasis involvement†‡ |
| NoLNresected | Total number of lymph node resected |
| NoLNinvolved | Total number of lymph node involved† |
| MarginInvolved | Margin involved |
| SizeLN | Size of largest involved lymph node† |
| Histo | Histopathology† |
| HistoType | Histological subtype† |
| TumorSize | Tumor size‡ |
| VI | Vascular invasion† |
| NoFociVIFTC | Number of foci of vascular invasion for follicular thyroid cancer (FTC)† |
| CI | Capsular invasion† |
| m-ETE | Microscopic extrathyroidal extension† |
| Mutation | Mutation |
| MutationType | Mutation type† |
| gETE | Gross extrathyroidal extension and invading location†‡ |
| Treatment | Thyroidectomy | Thyroidectomy |
| ThyroidectomyType | Thyroidectomy type |
| CompResectSurg | Completeness of resection (surgeon) † |
| CompResectPath | Completeness of resection (pathologist) † |
| Relation | Disease information (*Histo and HistoType*) | TypeOf | To indicate the relationship between Histo and HistoType  Example:   * Histologic type: Minimally invasive follicular carcinoma   Minimal invasive (HistoType)  follicular (Histo) |
| Disease information (Mutation and MutationType) | TypeOf | To indicate the relationship between Mutation and MutationType  Example:   * BRAF (MutationType)  Mutation (Mutation) |
| Treatment (Thyroidectomy and ThyroidectomyType) | TypeOf | To indicate the relationship between Thyroidectomy and ThyroidectomyType  Example:   * Complete thyroidectomy (HS, NIM)   Complete (ThyroidectomyType)  thyroidectomy (Thyroidectomy) |
| Attribute | Disease information (LN, Meta, VI, CI, m-ETE, Mutation, and gETE)  Treatment (CompResectSurg and CompResectPath) | Status | To indicate the presence of an entity (including LN, Meta, VI, CI, m-ETE, Mutation, gETE, CompResectSurg, and CompResectPath). The status of the listed entity can be ‘positive’, ‘negative’, or ‘unknown’.  Examples:   * Angionvasion: Not identified (VI[negative]) * Surface: Focally involved (CompRescPath [negative]) * The resection margins are clear of tumour cells (CompResectPath [positive]) |

**Notes:**

† Item for American Thyroid Association 2009 risk stratification

‡ Item for 8th edition of the American Joint Committee on Cancer TNM staging system of thyroid cancer

Yellow highlights serve as examples of annotation content that should be annotated within the free-text clinical notes, and they are used to illustrate how entities/relations/attributes should be annotated in BRAT.

**Example of annotating clinical note in BRAT**

A screenshot of a computer

Description automatically generated

**Detailed schema of annotation**

This section describes about how entities, attributes and relations should be annotated. For demonstration, highlights serve as examples of annotation content that should be annotated within the free-text clinical notes, and they are used to illustrate how entities/relations/attributes should be annotated in BRAT.

**Entity and attributes**

The entities are categorized into three categories, namely 1) temporal information, 2) disease information and 3) treatment.

1. **Temporal information**
2. ‘*Date*’ refers to the date of the pathology report or operation record generated. In the clinical notes, the date in the field ‘*Date Requested*’ or “*OT Start on*” should be annotated. For example:
   * Date Requested: 29/01/2024
   * OT Start on: 11-Aug 2021 at 10:33 OT End on: 11-Aug 2021 at 13:15
3. ‘*Age*’ refers to the patient’s age at diagnosis as mentioned in the clinical note. Integer value without unit (e.g. ‘y’, ‘yr’, ‘yrs’, ‘year’, ‘years’) should be labelled. For example:
   * Age: 62 yr
   * IF/67Y
   * M/39Y
4. **Disease information**
5. *‘AJCC8’* refers to the 8th edition of AJCC cancer staging classification of a patient. The stage, *‘I’*, *‘II’*, *‘III’*, *‘IVA’*, and *‘IVB’* should be annotated for *‘AJCC8’*. For example:

* AJCC 8th stage I (T1N0M0)

1. *‘Tstage’* means the primary tumor stage of TNM cancer staging. Values *‘T0’*, *‘T1a’*, *‘T1b’*, *‘T2’*, *‘T3a’*, *‘T3b’*, *‘T4a’*, and *‘T4b’* should be annotated. For example:

* AJCC 8th stage I (T1N0M0)
* pT1a: Tumor < or = 1 cm, limited to the thyroid
* Primary tumour staging: pT4a: Tumour more than 4 cm limited to thyroid

1. *‘Nstage’* refers to the regional lymph node stage of TNM cancer staging. Values *‘N0’*, *‘N1a’*, and *‘N1b’* should be labelled. For example:

* AJCC 8th stage I (T1N0M0)
* Stage pT4aN0 (pTNM, AJCC 8th edition)
* pN1a: Metastasis to reginal nodes

1. *‘Mstage’* indicates the distant metastasis stage of TNM cancer staging. Values *‘M0’* and *‘M1’* should be annotated. For example:

* AJCC 8th stage I (T1N0M0)
* pT2 NX M1 (right femur)
* Final TNM: pT2N0M0

1. *‘LN’* means any involvement and invading location of metastatic lymph node in addition to *‘Nstage’*. Both entity tags and attributes should be used to indicate the presence of lymph node metastasis. An attribute ‘Status’ could be positive (e.g. ‘present’, ‘identified’, ‘involved’, ‘noted’, ‘invading’, invading location (e.g. ‘central compartment’), level (e.g. ‘II’, ‘III’, ‘IV’ ‘V’, ‘VI’)), negative (e.g. ‘not identified’, ‘uninvolved’, ‘absent’) or unknown (e.g. ‘indeterminate’). For example:

* Metastasis to regional lymph node [attribute: positive]
* Lymph nodes are involved by metastatic carcinoma [attribute: positive]
* No nodes submitted or found [attribute: positive]
* Metastasis to unilateral, bilateral, or contralateral lateral neck lymph nodes (levels I, II, III, IV, or V) or retropharyngeal lymph nodes [attribute: positive]
* No evidence of malignancy [attribute: negative]
* One benign lymph node is presented. [attribute: negative]

1. *‘Meta’* refers to any involvement of distant metastasis in addition to *‘Mstage’*. Both entity tags and attributes should be used to indicate the presence of distant metastasis. An attribute ‘Status’ could be *‘positive’* (e.g. ‘present’, ‘identified’, ‘involved’, ‘noted’, ‘invading’, invading location (e.g. ‘lung’, ‘bone’, ‘liver’, ‘brain’), ‘metastatic’), *‘negative’* (e.g. ‘not identified’, ‘uninvolved’, ‘absent’) or *‘unknown’* (e.g. ‘indeterminate’). For example:

* Distant metastasis: Not applicable [attribute: negative]
* Metastasis to lung [attribute: positive]
* Site(s) of metastasis: Right femur [attribute: positive]

1. *‘NoLNresected’* identifies the total number of lymph node resected during operation or examination. Any cardinal numbers in numeric or words without unit should be labelled. If the total number of lymph node resected is not available, then the number of lymph node resected in each anatomical position should be annotated. For example:

* Metastatic papillary thyroid carcinoma (2/19)
* Two lymph nodes are sampled adjacent to the tumour
* Number of lymph nodes sampled around the thyroid gland: 21

1. *‘NoLNinvolved’* identifies the total number of lymph node involved. Any cardinal numbers in numeric or words without unit should be labelled. If the total number of lymph node involved is not available, then the number of lymph node involved in each anatomical position should be annotated. For example:

* Metastatic papillary thyroid carcinoma (2/19)
* Metastatic papillary carcinoma in one lymph node
* Number of lymph nodes involved: 1 (the larger piece)

1. *‘MarginInvolved’* refers to margins involved. The involved margin, including ‘anterior’, ‘posterior’, ‘lateral’, ‘tracheal’, ‘superior’, ‘inferior’, and/or ‘deep’ should be annotated. For example:

* The tumor focally involves the anterior margins (block 4).
* Surgical margins: Carcinoma reaches the anterior surface
* The anterior and posterior inked surfaces are focally involved.
* Margin status: R1, positive posterior margin

1. *‘SizeLN’* refers to the size of largest metastatic focus (or largest lymph node if the size of largest metastatic focus is not available). This could be in the form of 1-, 2- and 3-dimensional, such as ‘1mm’, ‘1mm x 2mm’ or ‘1mm x 2mm x 3mm’. The number and unit should be annotated together. If the information is contradicted between gross examination and microscopic examination, the one listed in gross examination should be annotated. For example:

* The largest focus is up to 0.6 mm.
* Size of largest metastatic deposit (mm), specify site: 0.6 mm
* Size of largest lymph node involved (mm): 10

1. *‘Histo’* indicates the histopathology of thyroid cancer. Examples of histopathology are ‘papillary’ and ‘follicular’. For example:

* Histologic type: Papillary carcinoma, classic subtype.
* Features are consistent with minimally invasive follicular carcinoma with capsular invasion only.

1. *‘HistoType’* indicates the subtype of histopathology. Examples of histological subtype are ‘classic’, ‘oncocytic’, ‘intrathyroidal’, ‘unifocal’, ‘multifocal’, ‘tall cell’, ‘columnar cell’, ‘hobnail variants’ (for papillary thyroid carcinoma [PTC]), ‘intrathyroidal, encapsulated follicular variant’ (for PTC), ‘infiltrative follicular’ (for PTC), ‘widely invasive’ (for follicular thyroid carcinoma [FTC]) ‘non-invasive’ (for FTC), ‘minimally invasive’ (for FTC), and ‘non-invasive follicular thyroid neoplasm with papillary like nuclear features [NIFTP]’. For example:

* Histologic type: Papillary carcinoma, classic subtype.
* Histologic type: Papillary carcinoma, infiltrative follicular subtype
* Involvement by PAPILLARY CARCINOMA, tall cell variant (4 mm)
* The overall findings are consistent with non-invasive follicular thyroid neoplasm with papillary-like nuclear features.

1. *‘TumorSize’* identifies the largest size of tumour of a patient. Similar to ‘SizeLN’, the data could be in the form of 1-, 2- and 3-dimensional and range (only under the condition where range is provided without specific values). The number and unit should be annotated together. For example:

* Right thyroid 1cm tumour with gross ETE to strap muscle
* Tumor size: Greatest dimension: 11 mm (macroscopic)
* Papillary carcinoma: 5 mm (block A5) and 1.5 mm (block A9)
* The major tumour nodule measures 2.5 cm in diameter and shows coarse calcification at the perphery.
* Papillary thyroid carcinoma (8 mm & 2 mm)
* On serial sections, a firm whitish tumor is noted at the upper pole, measuring 6 x 4 x 4 mm.
* The tumor measures 1.5 x 0.8 cm.
* Tumor > 2 cm, but < or = 4 cm, limited to thyroid
* TUMOR SIZE: Greatest Dimension: 2.5 cm

1. *‘VI’* refers to the status of vascular invasion. Both entity tags and attributes should be used to indicate the presence of vascular invasion. An attribute ‘Status’ could be ‘*positive’* (e.g. ‘present’, ‘involved’, ‘identified’, ‘invading’, ‘noted’, the invading location (e.g. ‘tissue’)), *‘negative’* (e.g. ‘not identified’, ‘uninvolved’, ‘absent’) or *‘unknown’* (e.g. ‘indeterminate’). For example:

* Angionvasion: Not identified [attribute: negative]
* Angiovasion: Absent [attribute: negative]
* Antioinvasion or lymphatic invasion is not identified [attribute: negative]
* Angionvasion: Extensive (= or > 4 vessels) [attribute: positive]
* VENOUS/LYMPHATIC (LARGE/SMALL VESSEL) INVASION (V/L): Present [attribute: positive]
* Lymph-Vascular Invasion: Not identified [attribute: negative]
* Lymphovascular invasion: not identified, but multiple positive lymph nodes are identified. [attribute: negative]
* Lymph-Vascular Invasion: Present and extensive (4 or more vessels) bilateral [attribute: positive]
* NO ANGIOLYMPHATIC INVASION OR EXTRATHYROIDAL EXTENSION. [attribute: negative]
* Blood Vessel Invasion: Not identified [attribute: negative]

1. *‘CI’* refer to the status of capsular invasion. Both entity tags and attributes should be used to indicate the presence of capsular invasion. An attribute ‘Status’ could be ‘*positive’* (e.g. ‘present’, ‘involved’, ‘identified’, ‘invading’, ‘noted’, the invading location (e.g. ‘tissue’)), *‘negative’* (e.g. ‘not identified’, ‘uninvolved’, ‘absent’) or *‘unknown’* (e.g. ‘indeterminate’). For example:

* Tumour capsule invasion: Present (focal) [attribute: positive]
* The tumor shows no capsule and the border is infiltrative. [attribute: negative]
* It has invaded beyond the capsule into the surrounding striated muscle. [attribute: positive]
* Multifocal capsular invasion is identified [attribute: positive]
* Extracapsular extension is present. [attribute: positive]
* TUMOR CAPSULAR INVASION: Present: Widely invasive [attribute: positive]
* Papillary thyroid carcinoma, forming a 0.8 cm tumor mass, confined within the thyroid capsule. [attribute: negative]

1. *‘m-ETE’* refer to the status of microscopic extrathyroidal extension. Both entity tags and attributes should be used to indicate the presence of microscopic extrathyroidal extension. An attribute ‘Status’ could be ‘*positive’* (e.g. ‘present’, ‘involved’, ‘identified’, ‘invading’, ‘noted’, the invading location (e.g. ‘tissue’)), *‘negative’* (e.g. ‘not identified’, ‘uninvolved’, ‘absent’) or *‘unknown’* (e.g. ‘indeterminate’). For example:

* Extrathyroidal extension is present [attribute: positive]
* Extrathyroidal extension: Absent [attribute: negative]
* Extrathyroidal extension: Not identified [attribute: negative]
* No definite microscopic extrathyroidal extension is seen. [attribute: negative]
* PAPILLARY THYROID CARCINOMA, MULTIFOCAL, INVOLVING BILATERAL LOBES, EXTENDING TO PERITHYROIDAL ADIPOSE TISSUE [attribute: positive]
* Extra thyroidal extension: positive for minimal extrathyroidal extension [attribute: positive]

1. *‘gETE’* refer to the invasion and invading location of gross extrathyroidal extension. Both entity tags and attributes should be used to indicate the presence of this clinical factor. An attribute ‘Status’ could be *‘positive’* (e.g. ‘present’, ‘involved’, ‘identified’, ‘invading’, ‘noted’, the invading location (e.g. ‘tissue’, “RLN”, “larynx”)), ‘negative’ (e.g. ‘not identified’, ‘uninvolved’, ‘absent’) or ‘unknown’ (e.g. ‘indeterminate’). For example:

* Right thyroid 1cm tumour with gross ETE to strap muscle [attribute: positive]
* Isthmic tumour, no ETE [attribute: negative]
* Gross extrathyroidal extension to overlying strap muscle, invasion to carotid sheath and wrapped around RLN at lower pole. [attribute: positive]
* 3cm right thyroid cancer with gross extra thyroidal extension to larynx and part of carotid sheath. [attribute: positive]
* and in part on the intraoperative finding of tumor invading outside the thyroid into the region of the recurrent laryngeal nerve. [attribute: positive]
* Extrathyroid Extension: Invades: perithyroidal fibradipose tissue and skeletal muscle. [attribute: positive]

1. *‘NoFociVIFTC’* indicate the number of foci of vascular invasion for FTC. These might be expressed in categorical or continuous form. For categorical form, the results could be no or minimal (e.g. ‘<4’, ‘< 4’, and ‘less than 4 vessels’) or extensive (e.g. ‘≥4’, ‘= or >4’, ‘= or > 4’, and ‘4 or more’). For continuous form, any integer number without unit should be labelled. For example:

* Angionvasion: Extensive (= or > 4 vessels)
* Angionvasion: Focal (<4 vessels)
* Lymph-Vascular Invasion: Present, extensive extent (4 or more vessels)
* LYMPH-VASCULAR INVASION: Present: Focal (less than 4 vessels)

1. *‘Mutation’* is the entity to classify whether the cancer mutation was identified. The *“Mutation’* test result could be *‘positive’*(e.g. ‘positive’, ‘identified’, ‘present’) or *‘negative’* (e.g. ‘negative’, ‘not identified’, ‘absent’, ‘no evidence’). An attribute will be annotated for ‘Mutation’ if the mutation status was presented in the clinical note. For example,

- On molecular studies, there is no evidence of BRAF, KRAS or NRAS hot spot mutation. [attribute: negative]

- BRAF V600 mutation IDENTIFIED at low level (see note). [attribute: positive]

- BRAF mutation POSITIVE (p.V600E, c.1799T>A) [attribute: positive]

1. *‘MutationType’* is the entity to indicate the type of the mutation. The options of *‘MutationType’* might include ‘BRAF’ (including ‘BRAF V600E’), ‘NRAS’ (including ‘NRAS codon 61’), ‘KRAS’ (including ‘KRAS codons 12/13’), ‘HRAS’ (including ‘HRAS codon 61’), ‘RET/PTC rearrangement’ (including ‘RET/PTC1’, ‘RET/PTC3’), ‘PAX8/PPARg rearrangement’, ‘TP53’, ‘NTRK1’ and ‘TERT promoter’. For example,

* BRAF V600E mutation IDENTIFIED at low level (see note).
* BRAF mutation POSITIVE (p.V600E, c.1799T>A)
* HRAS codon 61 mutation IDENTIFIED.

1. **Treatment**
2. ‘*Thyroidectomy*’ refers to the surgery, and the phrase ‘*thyroidectomy*’ should be annotated. For example:

* Complete thyroidectomy

1. ‘*ThyroidectomyType*’ refer to the type of thyroidectomy performed. Type of thyroidectomy include ‘*thyroid lobectomy*’ (left or right), ‘*hemithyroidectomy’* (left or right), ‘*isthmusectomy*’, ‘*open thyroid biopsy*’, ‘*partial*’, ‘*subtotal*’, ‘*total*’ and ‘*complete*’.

* Complete throidectomy
* Left hemithyroidectomy (HS, NIM (ETT+)
* Right hemithyroidectomy (HS, NIM)
* Received is a right hemithyroidectomy specimen measuring 4 cm from superior to inferior, 3.2 cm from medial to lateral and 1.2 cm from anterior to posterior.

1. ‘*CompResectSurg*’ and ‘*CompResectPath*’ indicate the completeness of resection of tumor in a procedure, from the perspective of surgeons and pathologists, respectively. The results could be *‘positive’* (e.g. ‘*complete*’, *‘yes’*, *‘clear’*, *‘touched’*), *‘negative’* (e.g. ‘*incomplete*’, ‘*partial*’, ‘*margin involved*’ in pathology report, ‘*gross tumour left behind*’ in operation record) or *‘unknown’*. An attribute ‘*Status*’ should be used to annotate the status of ‘*CompResectSurg*’ (in operation records) and ‘*CompResectPath*’ (in pathology reports). For example:

* Margin is clear. [attribute: positive]
* Surgical margins: uninvolved by carcinoma, closest 3mm from anterior margin [attribute: positive ]
* Cuff of muscle removed together with tumour, gross tumour clearance [attribute: positive]
* Focally, the tumor is separated from the surgical margin by thin fibrous tissue (0.6 mm). [attribute: positive]
* Resection margins: Focal involvement by invasive carcinoma at posterior surface. [attribute: negative]
* Margin status: Negative. [attribute: positive]
* Margin(s) involved by carcinoma [attribute: negative]

ALL MARGINS ARE FREE OF TUMOR [attribute: positive]

**Relation**

1. A relation of ‘*TypeOf*’ should be added from ‘*ThyroidectomyType*’ to ‘*Thyroidectomy*’. For example,

* Complete thyroidectomy (HS, NIM(ETT))

Complete  thyroidectomy

1. A relation of ‘*TypeOf*’ should be added from ‘*HistoType*’ to ‘*Histo*’. For example,

* Histologic type: Papillary carcinoma, infiltrative follicular subtype

infiltrative follicular  Papillary

* Histologic type: Papillary carcinoma, classic subtype

classic  Papillary

* Histologic type: Follicular carcinoma, widely invasive

widely invasive  Follicular

* Involvement by PAPILLARY CARCINOMA, tall cell variant (4 mm)

tall cell  PAPILLARY

1. A relation of ‘*TypeOf*’ should be added from ‘*MutationType*’ to ‘*Mutation*’. For example,

* BRAF V600E mutation IDENTIFIED at low level (see note).

BRAF  Mutation

* HRAS codon 61 mutation IDENTIFIED.

HRAS codon 61  Mutation