Melanoma for Surgical Pathology

# 1. Overview

Melanoma is a common and aggressive cancer that is associated with a high mortality rate if not treated promptly and appropriately. For operations and retrospective research, we need to identify melanoma patients and related diagnostic information, such as stage. However, this information is often missing, incomplete, or inconsistent in structured data. One reliable source for melanoma information is pathology reports, which play a critical role in melanoma diagnosis and treatment planning by documenting essential clinical and histopathological findings.

# 2. The NLP melanoma process

## 2.1. Cohort

Notes were selected from the Surgical Pathology domain, using the tables 'SurgPathSupplementDescript', 'SurgPathDiagnosis', 'SurgPathMicroscopicExam', and 'SurgPathPostOpDiagnosis'.

Additionally, the current system joins on the Surgical Pathology Morphology and Topography tables and requires the keyterms: a) ‘melanoma’, ‘lentigo maligna’, ‘junctional nevus’, or ‘melanotic’ for morphology, or b) ‘malignan’, ’tumor’, or ’neoplasm’ for morphology and ‘skin’ or ‘dermis’ for topography.

These tables are combined into a view containing unique identifiers for NLP processing.

## 2.2. NLP

We developed a rule-based NLP system using the medSpaCy framework, which contains common rules and packages specifically created for medical text. This system employs a series of sequential pipeline components for matching and relating relevant concepts. In order, the melanoma pipeline: 1) chunks text into smaller units (i.e. tokenization), 2) tags relevant terms, such as measurement values, histology terms, and anatomic sites, 3) uses combinations of tagged terms to mark relevant histologies (i.e. Named Entity Recognition), 4) uses combinations of tagged terms to mark related concepts within a dynamic scope and direction of the histologic entities, such as topography, depth, temporal, negation, and uncertainty information, 5) extracts entities and related features and parses output for measurement values, 6) aggregates all histologies and associated features at a biopsy sample level, and incorporates sample level logic for combining and interpreting results, and 7) aggregates information for document-level analysis.

## 2.3. The NLP melanoma output tables

Tables are currently stored within VINCI research spaces.

## 2.4. Interpretation of output tables and limitations

In addition to the data dictionary, the following is an overview of key decision points that affect interpretation of the output.

**ANCHORED ON DIAGNOSIS -** The NLP anchors all matches and relationships on a diagnosis. Diagnoses include any described melanoma or melanoma-related histology, some common non-melanoma cancer types (such as basal/squamous cell carcinoma), and some frequent skin-related non-cancerous cases, such as nevus or scar tissue.It should be noted that:

* Non-melanoma cases are used by the NLP pipeline as exclusion criteria and are NOT validated nor the focus of this project. These tables are intended to be used for melanoma/melanoma in-situ/lentigo maligna.
* The other measurements and features must be within a scope of a diagnosis (generally this is around 80 'tokens'). If a skin-related condition is NOT captured or the feature is outside of the scope, the feature will NOT be captured. Topography matches will also end the scope.

**OVERLAPPING DIAGNOSIS -** It is not uncommon for multiple diagnoses to be associated with a topography. In these cases, multiple flags (1's) will appear for the various diagnoses. For instance, the sample may contain melanoma in-situ on margins, but a different histology for the invasive component. Some of these cases could become errors with the sample-level rollup.

**DIAGNOSIS CATEGORIZATION -** The current categorization of melanoma types is as follows:

* Cutaneous: desmoplastic, epithelioid, nevoid, nodular, melanoma ‘not otherwise specified’, superficial spreading, spindle cell, spitzoid, letiginous, acral, amelanotic, hutchinsons melanotic freckle, balloon cell
* Mucosal: mucosal, melanoma with mucosal-related topography
* Ocular: choroidal, ocular, melanoma with an eye topography
* In-situ/lentigo maligna: These are separate columns with their own flags, and generally do not roll up to the other categories

There are additional edge cases for categorization, including the following:

* There are some cases that may only list a 'melanoma\_unspecified' (ie. no captured histology), or use a 'melanoma\_unspecified' term with reference the more specific histology. In these cases it will almost always flag as 'cutaneous' if there is also no lentigo maligna nor in-situ histology in the rollup (presumably in these cases the unspecified melanoma refers to in-situ).
* If there is a Breslow depth with an unspecific melanoma or lentigo maligna, it will also be flagged as cutaneous, even if in-situ is also flagged.

**FEATURE SPECIFICITY** - Some features may only attach to melanoma diagnoses, while others may attach to all diagnoses. This depends on whether the diagnosis is used to prevent some reasoning errors. As an artifact of this system, some features may show up for non-melanoma samples, but it is not guaranteed. This is another reason the focus should be on samples containing a melanoma diagnosis, which will consistently match the features.

**TOPOGRAPHY LEVEL ROLLUP** - This system aggregates all information within a sample. This rollup depends on a Topography match, which must fall within the scope of 200 'tokens' from a diagnosis. All diagnoses (within 200 tokens of topography) and all diagnosis features (which falls within usually ~80 tokens of the diagnosis) will be aggregated into the same row. The following should be noted:

* All diagnoses that do not have a topography match are rolled up together into a single row (per document) with no listed topography (ie. topography is empty).
* Because of the dependency of correct topography matching, it is possible for:
  + information for a sample to be incorrectly split into multiple rows (eg. there's a topography match part-way through the information of a sample, or a topography's scope ends too soon)
  + information from multiple samples to get mixed (eg. there's missing topography matches, so multiple sample information get rolled up into an empty topography)

**NOTABLE ROLLUP DECISION POINTS -** There are several points of reasoning throughout each point of the roll up processes. Some of these required clinical input, and others were used to eliminate erroneous matches.

Here's a summary of the different elements for each stage of the rollup: NLP matches -> feature values -> diagnoses -> sample -> document -> document staging -> patient

* For NLP matches -> feature values:
  + Breslow depth value will only be extracted if there is only one distinct value contained in the matched span. This accounts for some lacking Breslow depths when two values are close together near depth keyterms.
* For features -> diagnosis:
  + Context modifiers, such as temporality or negation, update diagnosis and metastasis status, but not other associated features
* For Diagnoses -> sample:
  + Features for the diagnosis is set as the highest severity when multiples of the same feature are contained:
    - For measurements, such as depth: the greatest values are kept
    - Presence is ranked above absence of other features, such as ulceration, metastasis, and histology
* For Sample -> document (for internal use):
  + Data is filtered for samples containing cutaneous melanoma, ocular melanoma, mucosal melanoma, melanoma in-situ, lentigo maligna, and historical melanoma positive cases, then aggregated on a document level. This eliminates any potential mixing of non-melanoma information with melanoma samples in each document.
  + A small percentage of documents have multiple samples with different melanoma values. These cases will mix the most severe of each of the features of the samples.
  + Only explicit ulceration presence is considered ulceration ('Impending' is not)
  + Document table is joined to include additional structured information
* For document -> staging (for internal use):
  + Breslow depth, ulceration status, cutaneous and in-situ status, and metastasis presence and location are used to determine TNM and clinical stage group
  + Documents mentioning metastasis without a topography are 'MX' assigned the nonexistent clinical stage '-1'