Source: https://huggingface.co/datasets/histai/HISTAI-metadata
File-type: .tiff
Compatible with: openslide, tifflib

# HISTAI Dataset

HISTAI is a comprehensive whole-slide image (WSI) pathological dataset spanning multiple medical specializations. Slides are anonymized and organized into specialized subsets by organ systems or pathology types.

If you wish to support, sponsor, or obtain a commercial license for HISTAI data, please contact us at models@hist.ai.

For details refer to our report:

- HISTAI: An Open-Source, Large-Scale Whole Slide Image Dataset for Computational Pathology

This repository contains metadata and references to images hosted separately. Individual slide images are accessible from specialized Hugging Face datasets.

---

## Dataset Structure

Slides are stored across specialized datasets hosted on Hugging Face. Each specialized dataset contains anonymized slides organized by cases:

histai/<dataset_name>/case_<case_id>/slide_<stain>_<slide_number>.tiff

or

histai/<dataset_name>/case_<case_id>/slide_<magnification>_<stain>_<slide_number>.tiff

Most of the slides are stained with Hematoxylin and Eosin (H&E) and scanned at 20X magnification. If a slide differs in magnification from 20X, this information is embedded in the slide filename, as shown above.

Currently available specialized datasets:

- HISTAI-hematologic
- HISTAI-gastrointestinal
- HISTAI-breast
- HISTAI-thorax
- HISTAI-skin-b2
- HISTAI-skin-b1
- HISTAI-colorectal-b1
- HISTAI-colorectal-b2
- HISTAI-mixed

## Metadata

The master repository includes comprehensive metadata in JSON format for each slide/case, containing detailed pathological, clinical, and technical information:

| Field | Description | Example |
|---|---|---|
| `diagnosis` | Incoming clinical notes | Benign skin neoplasms. |
| `conclusion` | Final pathological conclusion | Intradermal melanocytic nevus of the skin. |
| `diff_diagnosis` | Differential diagnostic notes (if available) | |
| `micro_protocol` | Microscopic description | Skin: Intradermal melanocytic nevus of the skin. Microscopic description: ... |
| `additional_info` | Any additional clinical/pathological notes | "A repeat review of the histological specimens was performed, including ... |
| `age` | Patient age (years) | 37 |
| `gender` | Patient gender | f |
| `icd10` | ICD-10 classification | D22 |

| | | |
|---|---|---|
| `specialization` | Medical specialization or organ system | Skin |
| `case_mapping` | Reference to slide images | histai/HISTAI-skin-b2/case_13384 |
| `grossing` | Gross examination details | "Head and neck: One fragment, measuring 2×4 mm, gray, firm, with ... |

---

## Statistics

| Dataset | Total Slides | Total Cases |
|---|---|---|
| histai/HISTAI-hematologic | 214 | 214 |
| histai/HISTAI-gastrointestinal | 202 | 120 |
| histai/HISTAI-breast | 1,925 | 1,692 |
| histai/HISTAI-thorax | 829 | 657 |
| histai/HISTAI-skin-b2 | 43,757 | 20,621 |
| histai/HISTAI-skin-b1 | 7,710 | 1,778 |
| histai/HISTAI-colorectal-b1 | 5,379 | 998 |
| histai/HISTAI-colorectal-b2 | 94 | 62 |
| histai/HISTAI-mixed | 52,691 | 21,137 |

- **Total slides**: 112,801
- **Total cases**: 47,279
- **Slides at x40 magnification**: 2,463
- **Slides at x20 magnifications**: 110,338
- **H&E slides**: 92,536
- **IHC slides**: 16,920
- **Other stains**: 3,345

Example of IHC Entry:

age: 76

gender: f

icd10: C43.7

specialization: Skin

case_mapping: histai/HISTAI-skin-b2/case_04631

diagnosis: Melanoma of the skin of the right lower extremity.

conclusion: Superficial spreading melanoma of the thigh, pT4b. Removed within clear margins. PD-L1 expression results (clone 28-8): PD-L1 expression < 1%. PD-L1 expression results (clone 22C3): MEL Score 2: >/=1<10%, TPS 2%, CPS = 1. Membranous PD-L1 expression (SP142) in tumor and tumor-infiltrating inflammatory cells is absent. Currently, there are no guidelines for interpreting IHC results with the PD-L1(SP142) clone in skin melanomas. Membranous PD-L1 expression (SP263) is detected in 2-5% of tumor cells. Membranous PD-L1 expression (SP263) is detected in 20% of tumor-infiltrating inflammatory cells. Currently, there are no guidelines for interpreting IHC results with the PD-L1(SP263) clone in skin melanomas.

diff_diagnostic:

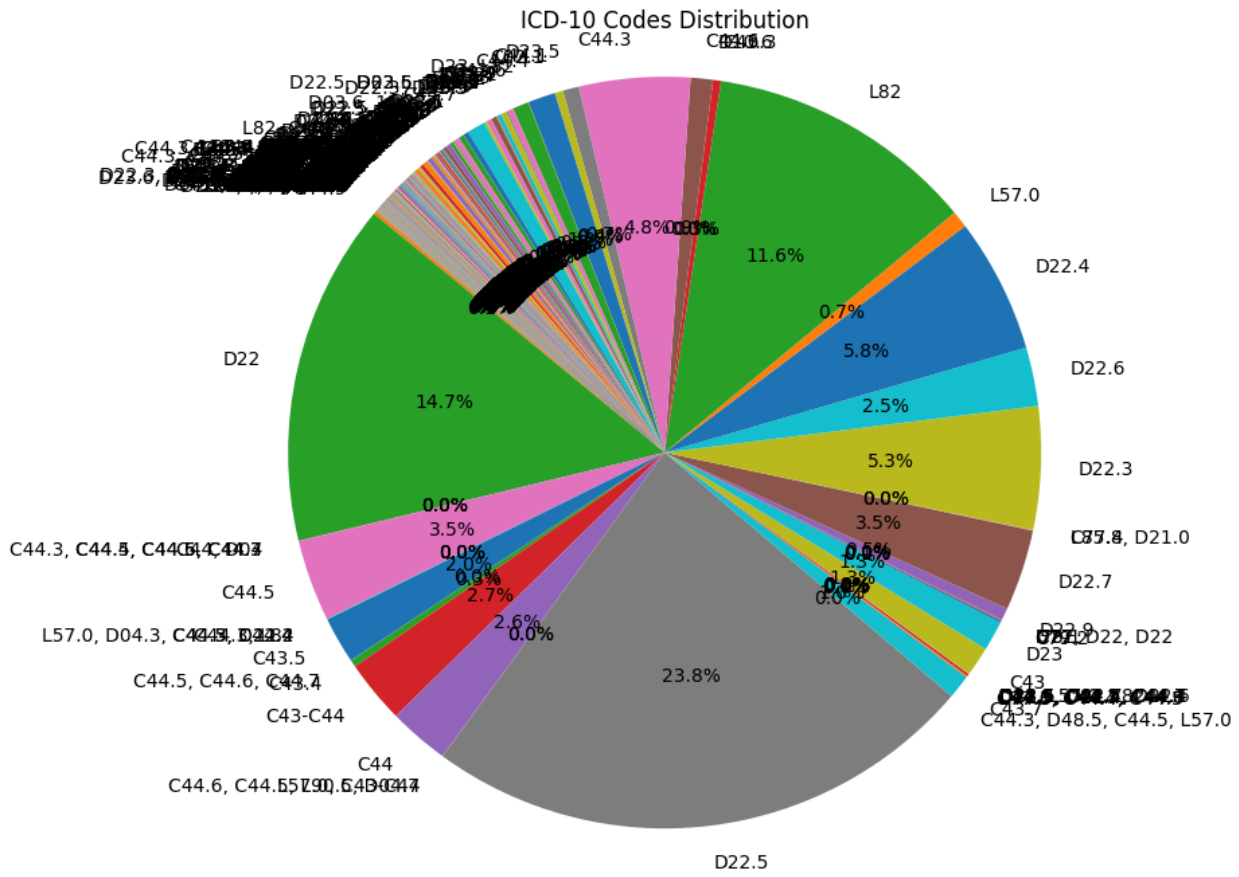
grossing: Skin: Skin flap. Marking: 1. Area from which the skin flap was taken: right thigh. Skin flap size (mm): 72×55×26. Presence of subcutaneous fat, dermis: with subcutaneous fat, with dermis. Subcutaneous fat thickness (mm): 24. Dermis thickness (mm): 2. Lesion localization: on the surface. Surface formation: nodule, spot. Lesion size (mm): spot 20×15 with a central nodule 14×14×4. Lesion color: black, dark brown, heterogeneous, gray. Border: clear. Surface formation: finely bumpy, with erosion. Distance to lateral resection margin (mm): 14. Margin marking with ink: present. Comment: Block marking: 1-2; 3-4; 5-6 - 1 slice of the lesion; 7-lesion; 8-9 - lesion periphery; 10-11- deep margin; 12-13- lateral margin.

micro_protocol: Skin: Melanoma. Operation: excision. Tumor localization: skin of the posterior surface of the right thigh. Tumor size: largest dimension (cm): 2. Macroscopically identifiable satellite nodules: not identified. Histological type: invasive melanoma. Invasive melanoma: superficial spreading melanoma. Maximum tumor thickness (Breslow): precise maximum thickness (mm): 5.8. Ulceration: present. Skin regression: Regression: not identified. Microsatellites: not identified. Resection margins: peripheral margin, deep margin. Peripheral margin: clear of invasive melanoma. Distance from invasive melanoma to the peripheral resection margin (mm): 14. Deep margin: clear of invasive melanoma. Distance from invasive

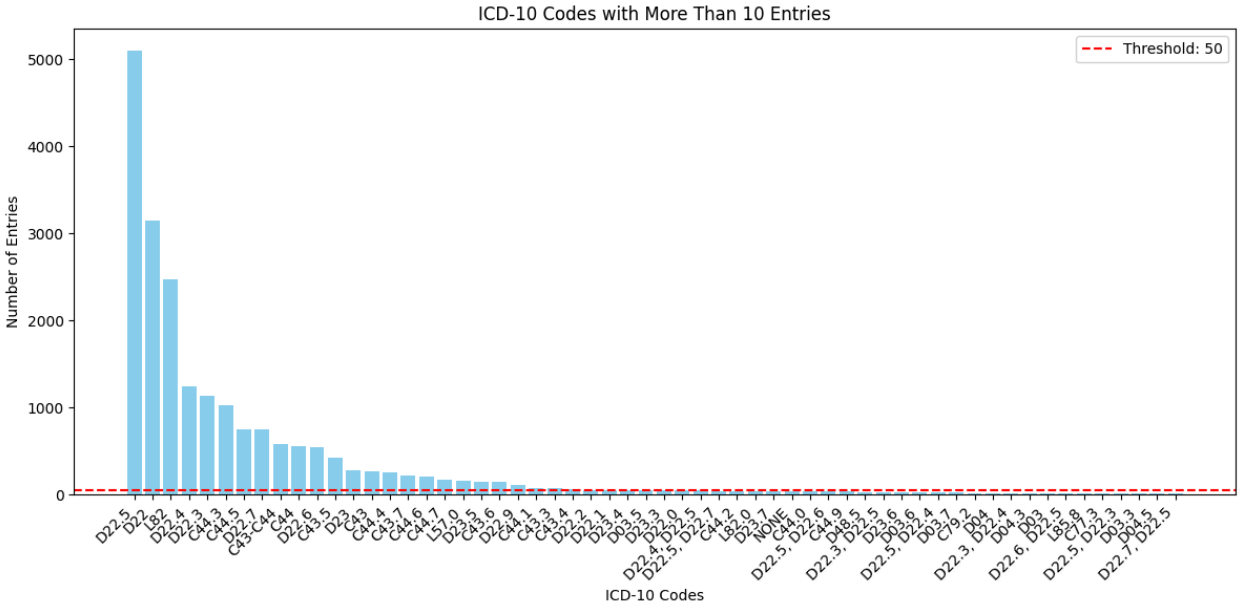
melanoma to the deep resection margin (mm): 20. Mitotic count: Mitotic count/mm2: 8. Anatomical level (Clark): IV (melanoma invades the reticular dermis, melanoma at the level of skin appendages). Lymphovascular invasion: not identified. Neurotropism: not identified. Tumor-infiltrating lymphocytes: present, mildly expressed. Tumor growth phase: vertical. TNM staging classification (AJCC 8th Edition) Primary tumor (pT): pT4: Melanoma > 4.0 mm thick, ulceration status unknown or not specified; pT4b: Melanoma thicker than 4.0 mm with ulceration. Additional studies: PD-L1. PD-L1 expression study performed: clone 28-8. Control tissue staining results: control passed. Sufficient tumor tissue: yes. Sample type: melanoma. PD-L1 expression result (clone 28-8, melanoma): Number of positively stained tumor cells: 0; PD-L1 expression < 1%. Guideline: PD-L1 IHC 28-8 pharmDx Melanoma Interpretation Manual, 19 January 2016/07 July 2016 (Agilent). Additional studies: PD-L1. PD-L1 expression study performed: clone 22C3. Control tissue staining results: control passed. Sufficient tumor tissue: yes. Sample type: melanoma. PD-L1 expression result (clone 22C3, melanoma) MEL Score 2: >/=1<10%. Guideline: PD-L1 IHC 22C3 pharmDx Interpretation Manual – Melanoma, 22 november 2017 (Agilent). Additional studies: PD-L1. PD-L1 expression study performed: clone SP142. Control tissue staining results: control passed. Sufficient tumor tissue: yes. Sample type: melanoma. For evaluation of the expression of this marker, it is mandatory to consult the following guideline https://diagnostics.roche.com/content/dam/diagnostics/us/en/products/v/ventana-pd-l1-sp263-assay/PD-L1-SP263-Bladder-Cancer-Sell-Sheet.pdf: help-content. Comment: Expression in tumor cells and tumor-infiltrating inflammatory cells is absent. Additional studies: PD-L1. PD-L1 expression study performed: another clone; sp 263. Control tissue staining results: control passed. Sufficient tumor tissue: yes. Sample type: melanoma. For evaluation of the expression of this marker, it is mandatory to consult the following guideline https://diagnostics.roche.com/content/dam/diagnostics/us/en/products/v/ventana-pd-l1-sp263-assay/PD-L1-SP263-Bladder-Cancer-Sell-Sheet.pdf: help-content. Comment: Membranous PD-L1 expression (SP263) is detected in 5-10% of tumor cells. In 20% of tumor-infiltrating inflammatory cells, membranous PD-L1 expression (SP263) is detected.

additional_info:

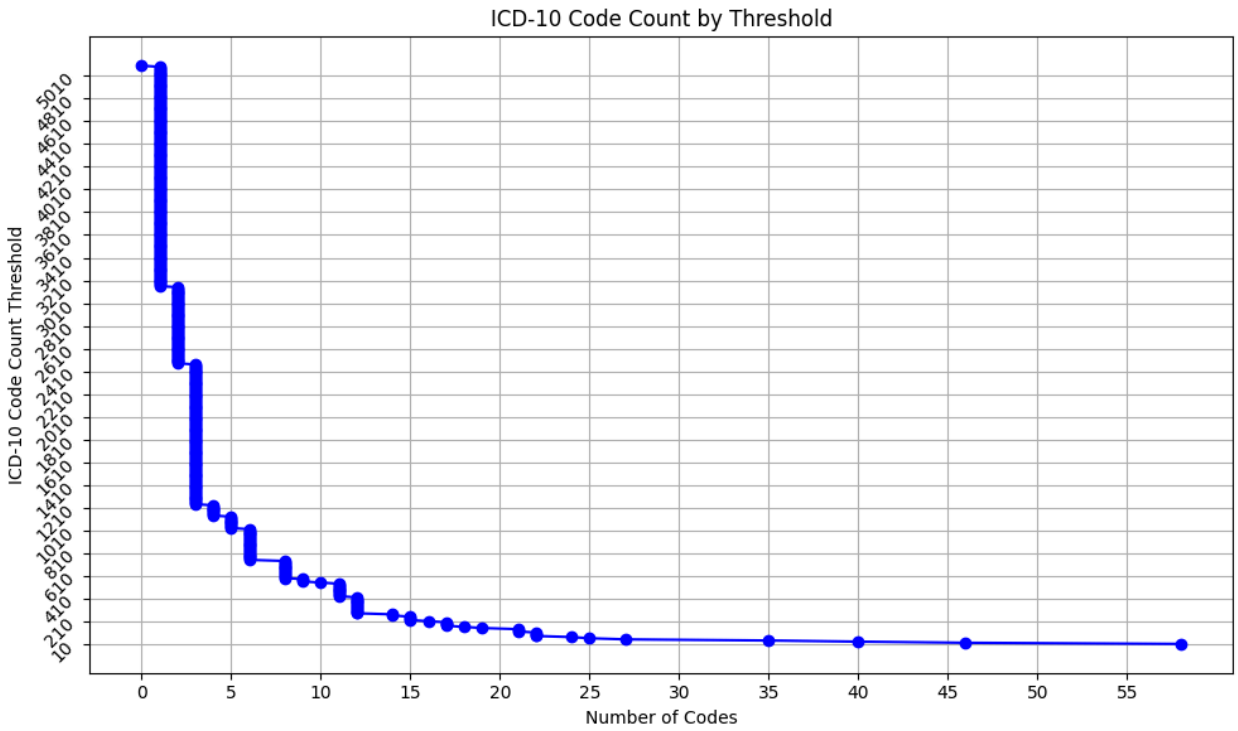
Haut Fälle sind verteilt über "specialization"="Skin" und "unknown". Insgesamt 21677, nach Filter von "IHC" in "conclusion" und "micro_protocol" noch 21456. Diese enthalten 401 verschiedene ICD-10 Code Combinationen. Durchschnittlich enthält ein Fall 2.3 WSI.




Dabei sind es 174 einzigartige ICD-10 Codes mit durchschnittlich 49.72 Fällen pro Code und eine Std von 266.34. Nach einem Threshold mit mind 10 Cases pro Code gibt es diese Verteilung. Rote Linie bei 50 Cases.




Daraus ergibt sich dieser Verlauf.




Es gibt vielfach fragwürdige/falsche ICD10 zu Diagnosebezeichnungen;
Beispiel: Seborrhoische Keratose ICD-10 L82 nach https://icd.kbv.de wird entweder ohne ICD-10 Bezeichnung versehen, oder D22/D23/D22.4/C44.3

Entry 21321 contains 'Seborrhoeic keratosis':
  age: 36.0
  gender: m
  icd10:
  specialization:
  case_mapping: histai/HISTAI-skin-b2/case_00224
  diagnosis: Seborrheic keratosis.
  conclusion: Seborrheic keratosis of the abdominal skin, mixed (papillomatous and keratotic) histological variant, removed with clear margins.
  diff_diagnostic:
  grossing:
  micro_protocol: Skin lesion with papillomatous proliferation of basaloid cells, marked hyperkeratosis, and horn cysts within the epithelium. Melanin pigment deposition is present in the basal layer of the epidermis. A thin strip of underlying tissue is seen, with no elements of the lesion along the resection line.
  additional_info:

Entry 21345 contains 'Seborrhoeic keratosis':
  age: 67.0
  gender: m
  icd10: D23

specialization:
case_mapping: histai/HISTAI-skin-b2/case_15905
diagnosis: Seborrheic keratosis.
conclusion: Seborrheic keratosis of the skin.
diff_diagnostic:
grossing:
micro_protocol:
additional_info: The morphological picture corresponds to seborrheic keratosis. Removed within the limits of healthy tissues.

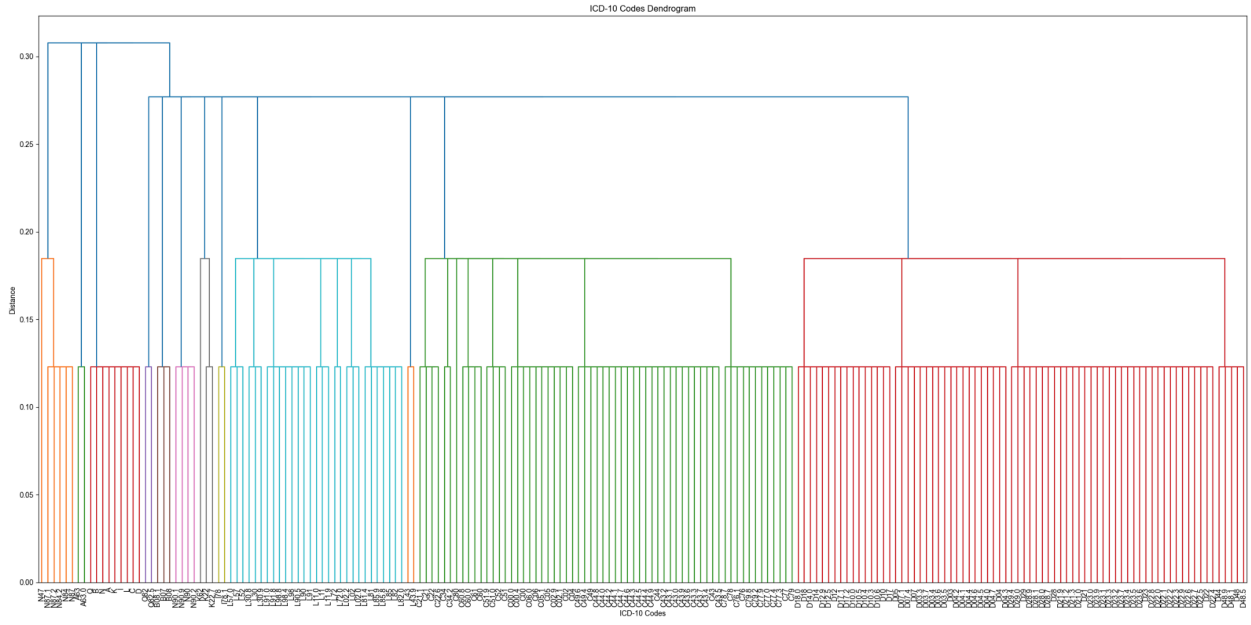
Entry 21348 contains 'Seborrhoeic keratosis':
age: 37.0
gender: f
icd10: D22
specialization:
case_mapping: histai/HISTAI-skin-b2/case_17753
diagnosis: Seborrheic keratosis? Papillomatous nevus?
conclusion: Seborrheic keratosis of the skin of the left breast.
diff_diagnostic:
grossing:
micro_protocol:
additional_info: Papillary lesion lined with multilayered epithelium without atypia, with pronounced hyperkeratosis and parakeratosis forming keratinous cysts.

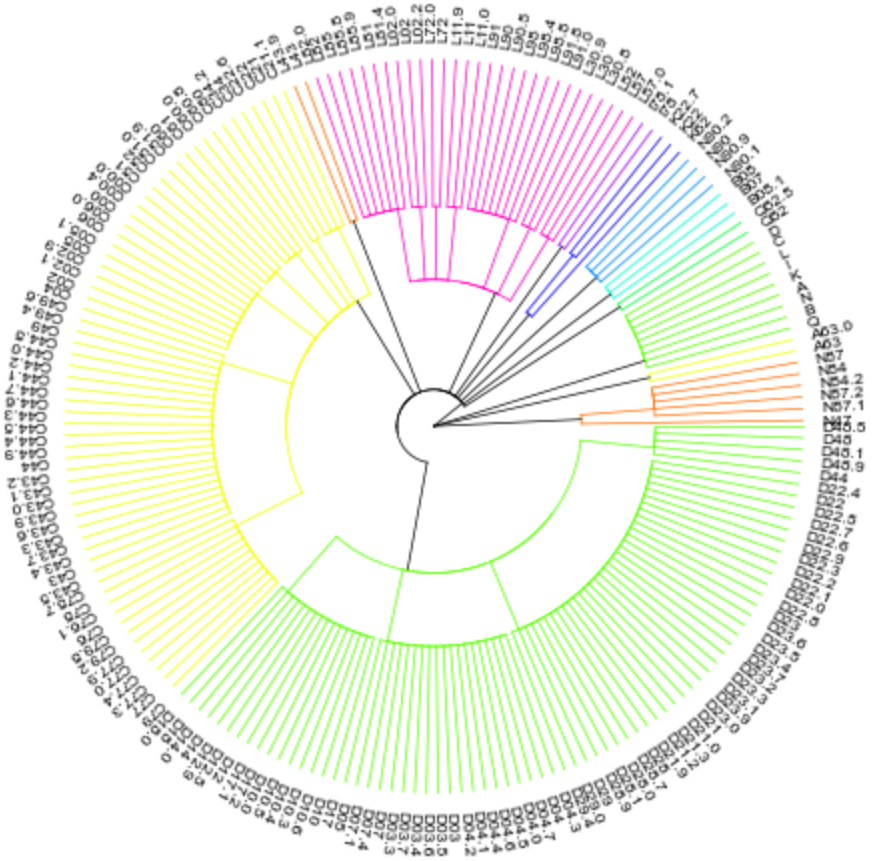
Entry 21378 contains 'Seborrhoeic keratosis':
age: 41.0
gender: f
icd10: D22.4
specialization:
case_mapping: histai/HISTAI-skin-b2/case_17296
diagnosis: Seborrheic keratosis, recurrence.
conclusion: Seborrheic keratosis of the skin. Removed within healthy tissues.
diff_diagnostic: The histological picture corresponds to seborrheic keratosis of the skin.
grossing:
micro_protocol:
additional_info: Skin with hyperkeratosis, acanthosis, formation of keratinous cysts, and hyperpigmentation of the basal epidermal layer. Removed within healthy tissues.

Entry 21437 contains 'Seborrhoeic keratosis':
age: 72.0
gender: f
icd10: C44.3
specialization:
case_mapping: histai/HISTAI-skin-b2/case_02043

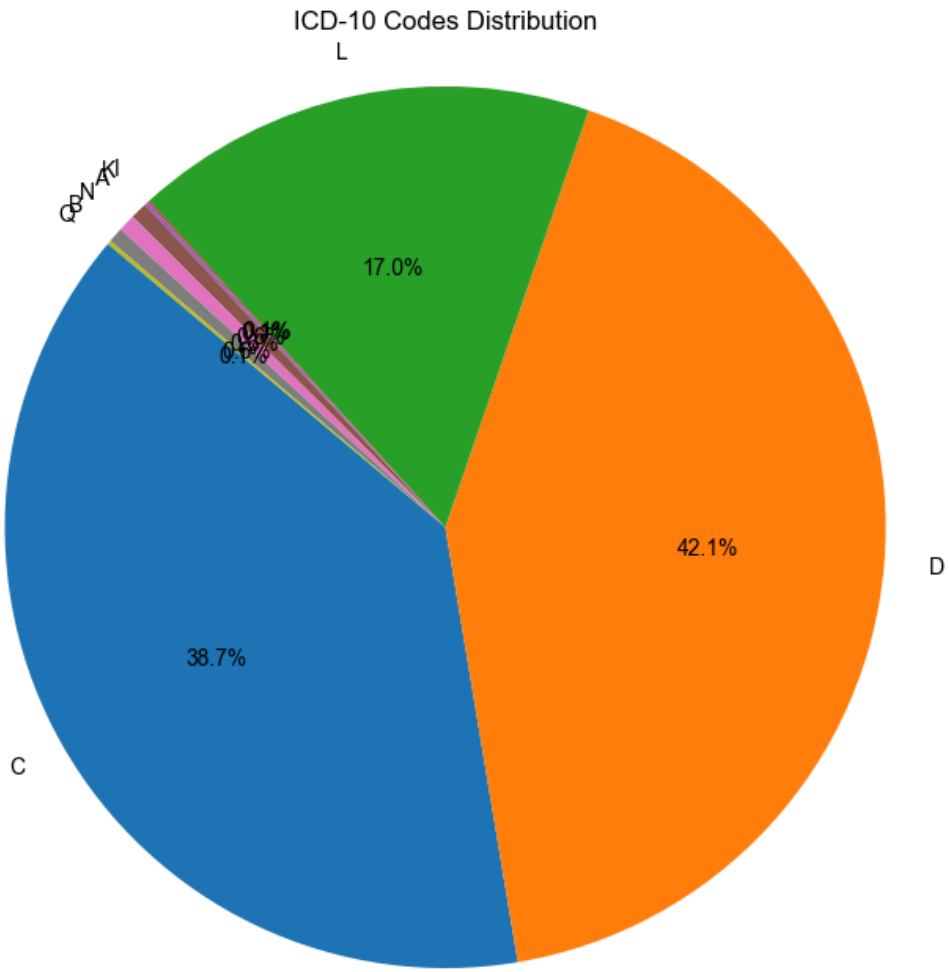
diagnosis: Seborrheic keratoma? Skin horn? Papilloma?
conclusion: Common skin wart. Removed within the limits of healthy tissue.
diff_diagnostic: The histological picture corresponds to a common skin wart.
grossing:
micro_protocol:
additional_info: Skin with papillomatosis, pronounced hyper- and parakeratosis, and epidermal hypergranulosis. Removed within the limits of healthy tissue.

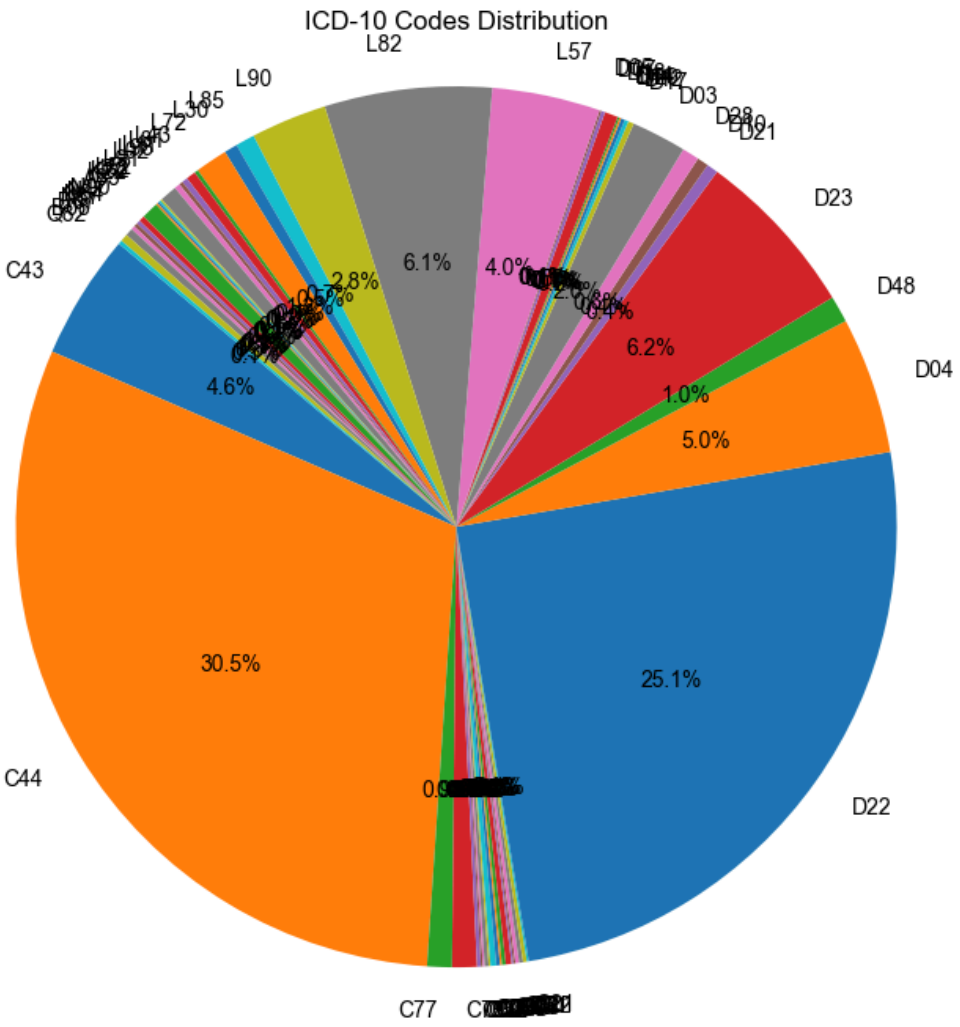
Als nächstes die Granularität der ICD-Coverage:

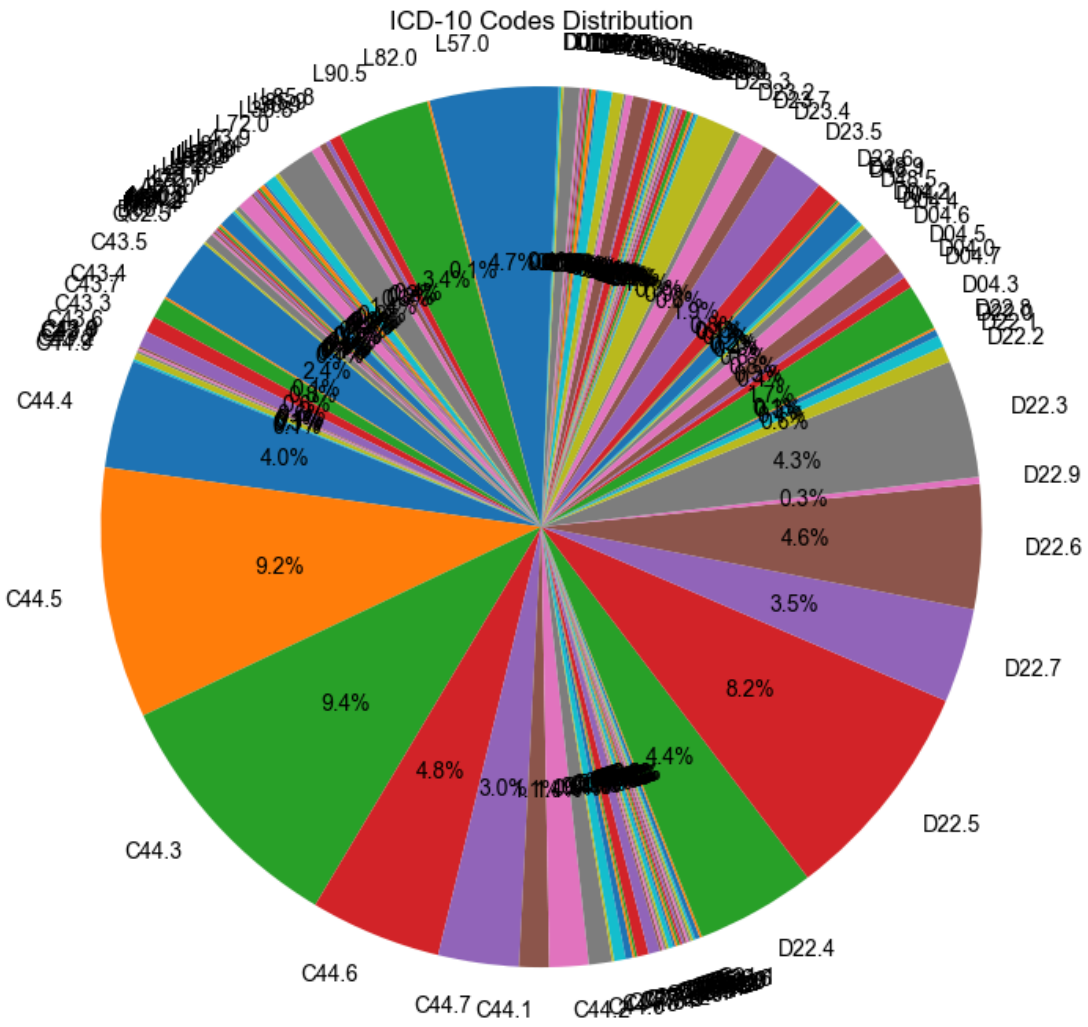




Und die Verteilung nach in Kapitelstufen:


ICD-10 Codes Distribution
L
17.0%
42.1%
D
38.7%
C


ICD-10 Codes Distribution
L82
L57
L90
L85
L30
L72
D03
D21
D23
D48
D04
C43
6.1%
4.0%
2.8%
6.2%
1.0%
5.0%
4.6%
30.5%
25.1%
C44
D22
C77


ICD-10 Codes Distribution
L57.0
L82.0
L90.5
L72.0
C43.5
C43.4
C43.7
C43.3
C44.4
C44.5
C44.3
C44.6
C44.7
C44.1
D22.4
D22.5
D22.7
D22.6
D22.9
D22.3
D22.2
D04.3
D04.7
D04.0
D04.5
D04.6
D23.5
D23.4
4.0%
9.2%
9.4%
4.8%
3.0%
4.4%
8.2%
3.5%
4.6%
0.3%
4.3%
3.4%
2.4%