

## Appendix S1

### Data Cleaning

Each retrieved pathology record from *Patobank* could contain descriptions of multiple tumours removed in the same clinical setting, and therefore an algorithm was created to split these records into single specimen results. Subsequently the inclusion and exclusion criteria which are described in the main manuscript had to be applied to each new single specimen record to ensure consistency. Macro- and microscopic free text descriptions were furthermore assessed to ensure exclusion of all excisions with intraoperative margin assessments (e.g. frozen sections, repeat sections), biopsies (punch or incisional), and curettages. A second round of filtering was then performed based on key words presented in Supplementary Table S1.

To extract the macroscopic specimen resection size, a search algorithm was developed. The algorithm searched each free-text macroscopic description for the largest dimension number in millimetres, while ignoring other nonrelevant numerical data, such as record ID and numbered location (often described by the pathologist using a 12-hour clock as a reference). The algorithms were all built in Python (version 3.7.4) using the following packages: Pandas (version 2.2.1) and Numpy (version 1.26.4).

### Defining Variables

The following data were extracted from records: patient's unique civil registration number (numerical), macroscopic specimen description (free text), microscopic specimen description (free text), and SNOMED CTs (encoded). These data were operationalised into the variables: sex, age, anatomic location, surgical specimen resection size, histopathologic tumour subtype, and margin outcomes.

Sex and age were deduced directly from the unique Danish civil registration numbers (i.e. an even last digit in the number = female). Anatomic location was extracted according to the SNOMED CT and aggregated into anatomic groupings by assessing the number of entries for each location and meaningfulness. Histopathologic subtype was extracted according to the SNOMED CT as well as from the free text descriptions. The latter was necessary due to some specimens not being subtype SNOMED CT encoded. Further, in specimens registered with a combination of two or more histopathologic subtypes, the most histopathologically aggressive subtype was chosen in the following descending order: basosquamous, morpheaform, infiltrative, micronodular, nodular, or superficial.

In specimens missing subtype SNOMED CT encoding, free text descriptions were used to extract subtypes, by algorithmically searching for the words nodular (and all variations of this), and not containing the words superficial, micronodular, infiltrative, etc. (and all variations of those). When these conditions were met, nodular became the registered subtype. When the search conditions were not met, 'not available' (N/A) became the recorded subtype. Further, fibroepithelial BCCs were included as nodular BCCs. A manual review was performed to check the accuracy of this algorithm.

Similarly, surgical specimen resection size was only available in free text in the macroscopic description of pathology records. Thus, an algorithm was built to read the largest noted specimen dimension in mm. A manual review of 100 random reads between 5-49 mm found only one algorithmic size determination error, and the error was deemed not systematic in nature. It was evident that records with surgical specimen resection sizes  $< 5$  mm were almost exclusively curettages and biopsies that had escaped previous data cleaning filters due to the sparse information in histopathology records of small lesions. This, in combination with the unlikelihood of proper surgical excisions smaller than 6 mm, we chose to exclude all excisions with sizes  $< 5$  mm and manually reviewed and excluded excisions with sizes between 5-6 mm.

*Supplementary Table S1: Exclusion Criteria Used in the Algorithm*

|                  | Exclusion performed based on Danish key words also including common, mistyped words                                                                                                       | Exclusion based on SNOMED CT                                                                                           |
|------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| Frozen section   | “Frys”, “frys”                                                                                                                                                                            | “P32200”, “PYYFRY”                                                                                                     |
| Biopsy           | “hudbiopsi”, “biopsi” & (-“exc”/“eks”), “tanse”, “ylinderform”, “tansebiopsi”, “2 væv”, “3 væv”, “vævsstykker”, “ragmenter”, “2 hudbeklædte vævsstykker”, “3 huder”, “2 huder”, “ragment” | “P30610”, “P30613”, “P3061A”                                                                                           |
| Curettage        | “flager”, “fnug”, “curet”, “Curet”, “urretage”, “uretage”, “urrettage”, “skrab”, “hudcurettage”, “ragmentere t hud”, “hud med fragmenter”, “udfragmenter”, “lager af hud”                 | “P30626”                                                                                                               |
| Relapsed tumour  |                                                                                                                                                                                           | “M80907”, “M80917”, “M80927”, “M80937”, “M80947”, “M80957”, “M80977”, “M80987”, “M809A7”, “M809B7”, “M809x7”, “ÆYYY07” |
| Repeat sections* |                                                                                                                                                                                           | “P30624”*                                                                                                              |

\* P30624 defined as repeat sections, i.e. re-resectates.

*Supplementary Table S2: Aggregation of SNOMED Clinical Terms for Anatomic Locations*

| Nose   | Forehead | Scalp  | Cheek  | Collum | Cheek  | Other face | Temple | Ear    | Lip    | Neck   | Chin   | Postauricular | Preauricular | Periorbital | Eye lid | Mandible | Eye canthus |
|--------|----------|--------|--------|--------|--------|------------|--------|--------|--------|--------|--------|---------------|--------------|-------------|---------|----------|-------------|
| T02140 | T02104   | T02102 | T0212D | T02300 | T0212C | T02120*    | T0211C | T0220B | T02151 | T02305 | T02155 | T02113        | T02114       | T0213G      | T0213E  | T0216A   | T02134      |
| T2114B | T01530   | T02100 |        |        | T02121 |            | T0211D | T0220A | T02152 | TY0610 |        | T0211H        | T0211L       | T0213H      | T0213F  | T0216B   | T0213M      |
| T21120 | TY0110   | TY0160 |        |        |        |            | T02111 | T02200 | T02150 |        |        | T0211G        | T0211K       | T0210N      | T02130  | T02160   | T0213N      |
| T2114A |          |        |        |        |        |            | TY015A | TXY10B |        |        |        | TY017B        | TY017D       | T0210M      | T02132  |          | T0213P      |
| TY0316 |          |        |        |        |        |            | TY015B | TXY10A |        |        |        | TY017A        | TY017C       | T02133      | TXX810  |          | T0213Q      |
| T21000 |          |        |        |        |        |            |        | TXY100 |        |        |        | TY0170        | TY013A       | T01520      | T0213C  |          |             |
|        |          |        |        |        |        |            |        | T0221X |        |        |        |               |              | T0152A      | TXX831  |          |             |
|        |          |        |        |        |        |            |        | T0221Y |        |        |        |               |              | T0152B      | T0213B  |          |             |
|        |          |        |        |        |        |            |        | TXY20B |        |        |        |               |              | T02105      | T0213A  |          |             |
|        |          |        |        |        |        |            |        | TXY200 |        |        |        |               |              |             | T0213D  |          |             |
|        |          |        |        |        |        |            |        | TXY20A |        |        |        |               |              |             | TXX832  |          |             |

\* T02120 is defined by SNOMED CT as “skin of face”.