# Overall structure of algorithm to identify cancer-stage based on data from the cancer-registry.

**History of algorithm:** Originally created by Erzsébet, later updated by Emese. The current version is created by Morten and Julie in June 2025. The structure has been updated, and the content/values have been revisited and updated by Julie for a project on dietary behavior among cancer survivors on the Better Health in Late Life (BHLL) cohort.

Disclaimer: We have focused on the Danish population aged 50-65 in the Autumn 2021, so we are not sure that we are covering all possible cancer-registration, but we expect to cover the big majority of cancer registrations. Also a few new values/combination might have been introduced after 2021.

## Structure:

1. Cancer type is identified. The staging is dependent on the cancer type, so the first step is to identify cancer type based on ICD10 codes (and in a few cases morphology codes as well). 21 common cancer types are selected for this staging algorithm.
2. Cancer registrations are split between the common (the 21 types mentioned above) and rare cancer types. We find 86% of the respondents with cancer to be within the common cancer types, and 14% to be within the rare types.
3. Among the common cancer types, registrations are split on whether they are classified based on TNM-codes or not. If they are not there are usually two reasons:
   1. Old registrations (registrations from before 2004 are not classified by TNM-codes).
   2. Three of the common cancer types are not classified by TNM-codes (Brain cancer, Hodgkin Lymphoma, Non-Hodgkin Lymphoma). Hodgkin Lymphoma, Non-Hodgkin Lymphoma are instead classified by Ann Arbor stage (C\_AA).
4. Focusing on Common cancer types that can be classified based on TNM-codes in the coming steps.
5. The TNM-codes from the cancer registry can have values like {AZCD13c, AZCD30, AZCD40…}. The translation for these codes can be found in the sks-browser. All the AZ-codes are downloaded from the SKS-browser and stored in the dataset WorkCanc.sks\_az\_koder. A format for translation of these codes is generated. The TNM codes are translated based on this format.
6. Now the TNM-values has values like {T1, T2b1, N1mic, M0}. These are simplified even further. Very often they can be aggregated down to the two-character values {T1, T2, T3, T4, N0, N1, N2, N3, M0, M3}. But there are exceptions for some of the cancer-types.
7. Next step is to use a table to translate the different combinations of T-, N- and M- values into stage. This is done by joining a big table (workdiet.cancer\_stage\_tnm\_20250606) with cancer-type specific translation of TNM codes into stage.
   1. The table build on cancer\_stage\_tnm\_20200810 from Emese. The main work on updating the algorithm by Julie, has been updating this table…
   2. A commented version of the table with Julies comments can be found as E:\ProjektDB\KEA\Workdata\708194\Morten Madsen\survey50+\Julie\_cancer\_diet\Output\cancer\_stage\_tnm\_apr2025\_JAS.csv (column: ‘set short\_stage to’)
      1. Besides Julies comments, a few redundant lines has been marked for deletion (by Morten) in: cancer\_stage\_tnm\_apr2025\_JAS\_MM.csv
   3. In the SAS\_program, it is possible to see which updates have been made between the different versions of the translation-table.
   4. In the table also appears values like “AnyT” meaning that the line is covering all possible values of T.
   5. And values like “Nx” or “Mx” is also possible - indicating that the given value Is not measured. This usually results in multiple possible stages.
8. In case of multiple lines matching, it is manually checked that these will result in the same stage.
9. A little postprocessing is done in the next step
   1. It is checked that all values have been translated by the table (if not manually updating of the algorithm/program is needed).
   2. A few cases that are not really cancer is deleted (e.g., Tis: in situ)
   3. If the M-value is Mx, the stage almost always has to be set to unknown (because the possibilities covers at least two very different stages).
      1. There is an exception to this for prostate-cancer, which is also implemented in this step.
10. The other cancer-types staged with Ann-Arbor (or c\_udbred).
11. All the cancer registrations are put together in one dataset.
12. If multiple cancer-registrations pr person exists, the most recent cancer type is picked (and previous\_cancers are set to mark that a previous cancer has occurred).

## Process by which the algorithm for assigning stage (stage\_short) based on TNM-codes was updated (Spring 2025):

* Information provided by Emese: The algorithm to define stage based on TMN was originally developed by external collaborator Frits Mulder and Erzsébet. In a subsequent study applying the algorithm external collaborator Vincent Lanting applied the algorithm with Emese. It is not entirely clear, but they may have used the AJCC (American Joint Committee on Cancer) classification system.
  + The AJCC staging is available on the webpage of the American Cancer society for each cancer type, e.g, [Prostate Cancer Stages | Staging of Prostate Cancer | American Cancer Society](https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/staging.html)
* The AJCC and UICC staging manuals are generally very similar, but UICC is sometimes more straight forward, e.g., for prostate cancer in which the AJCC also includes grade and PSA, or breast cancer which also includes HER2, grade, and hormone receptor status (and only examples are available on the webpage not a full staging manual)
  + The UICC manual 8th ed is available here: [ProQuest Ebook Central - Book Details](https://ebookcentral-proquest-com.ez.statsbiblioteket.dk/lib/asb/detail.action?pq-origsite=primo&docID=4792667)
* Julie check the TNM codes and the variable ‘stage\_stort’ against the UICC 8th ed – using clinical stage. Edits made to the algorithm are documented in: cancer\_stage\_tnm\_apr2025\_JAS\_MM.csv, column ‘Set stage\_short\_to’.
  + All the common 21 cancer types were checked, except:
    - Esophageal and billary cancers because the staging depends on subsites, which were not available to us, and because these cancer types are quite rare in the BHLL cohort.
    - Brain, Non-Hodgens and Hodgins lymphomas because TMN is not used for these cancers
  + For some codes the AJCC coding was also checked to make sure that edits made did not overwrite what was correct according to the AJCC coding.
  + The majority of stage\_short were in line with the UICC staging manual
  + If additional information to Tn, Nn, Mn available (e.g., T2a or T2b) in the algorithm but not in the staging manual, stage assigned according to Tn, Nn, Mn ignoring the postfix.
  + If the number is higher than what is mentioned in the UICC staging manual, it was treated as the highest number. E.g., N2 for Kidney treated as N1. (These are usually very rare)
  + Stage\_short set to Unknown: if more than one non-adjacent stages were possible, e.g. stage II and IV (this mostly occurred M was unknown, i.e., Mx)
  + Stage\_short set to X/X+1: if two adjacent stages were possible, e.g., II/III

## SAS-dataset / SAS-programs for BHLL studies:

The dataset **E:\ProjektDB\KEA\Workdata\708194\Morten Madsen\survey50+\Julie\_cancer\_diet\WorkDiet\prevalent\_cancer.sas7bdat**

Holds cancer stages (at the time of identification) for all Danes aged 50-65 in the Autumn 2021 and thereby includes the BHLL-cohort.

It has been created by the programs:

**E:\ProjektDB\KEA\Workdata\708194\Morten Madsen\survey50+\Julie\_cancer\_diet\05\_cancer\_type.sas**

And

**E:\ProjektDB\KEA\Workdata\708194\Morten Madsen\survey50+\Julie\_cancer\_diet\06\_cancer\_stage.sas**

These programs use the supporting dataset: **E:\ProjektDB\KEA\Workdata\708194\Morten Madsen\survey50+\Julie\_cancer\_diet\WorkDiet\cancer\_stage\_tnm\_20250613.sas7bdat**