
A health-system-scale, episode-resolved multimodal atlas of cancer cachexia

In the format provided by the
authors and unedited

Data Dictionary: De-identified Cachexia Episode-Level Table

File: deidentified_episodes_cachexia_20260821.csv. 52,236 patients; 58,817 rows (one row per patient with no detected episode, plus one row per detected weight-loss episode).

1. DMP_ID — De-identified patient identifier.
2. sex — Patient biological sex (Male/Female).
3. age_at_diag — Patient age at anchor/diagnosis date. Values >89 are coded as ">89".
4. CANCER_TYPE_DETAILED — Detailed cancer type label from MSK-IMPACT annotation. Cancer types with fewer than 20 patients in the cohort are censored and grouped together as "Other rare cancer types" (376 cancer types, 2,214 patients affected). All other cancer types are reported at full granularity.
5. ONCOTREE_CODE — OncoTree cancer type code. Censored to "OTHER" for the same rows grouped under "Other rare cancer types" in CANCER_TYPE_DETAILED.
6. STAGE_CDM_DERIVED — Clinical stage derived from the MSK cancer data management system.
7. bmi_cat_at_anchor_date — BMI category based on the first available BMI measurement on or after anchor date. Categories: Underweight (<18.5), Normal (18.5–24.9), Overweight (25.0–29.9), Obese (≥30).
8. episode — Binary flag indicating whether this row represents a detected cachexia episode (1) or a non-episode patient row (0).
9. episode_id — Episode number for this patient (1 = first episode, 2 = second episode, etc.). NULL for non-episode rows.
10. threshold — Detection threshold under which this episode was called. WL5_WL2BMI20 = 5% weight loss (or 2% if start_bmi_smooth < 20).
11. reached_WL10_6mo — Patient-level binary flag; whether the patient had any validated cachexia episode under the independently re-run 10% weight-loss (WL10) detection algorithm.
12. reached_WL15_6mo — Patient-level binary flag; whether the patient had any validated cachexia episode under the independently re-run 15% weight-loss (WL15) detection algorithm.
13. start_day — Days since anchor date at which the episode detection window was triggered.
14. end_day — Days since anchor date at which the episode ends. NOT the nadir.
15. episode_duration — Duration of the episode in days (end_day – start_day).
16. start_bmi_raw — Raw BMI value at start_day from the EHR.

17. `start_bmi_smooth` — EWMA-smoothed BMI ($\alpha=0.2$) at `start_day`. Used as the baseline for episode detection and for computing `weight_loss_max`.
18. `bmi_cat_at_episode_start` — BMI category based on `start_bmi_smooth`. NULL for non-episode rows.
19. `weight_loss_start_to_end` — Percent weight loss from episode start to episode end. May be less than `weight_loss_max` if partial BMI recovery occurred before `end_day`.
20. `nadir_day` — Days since anchor date at which the lowest smoothed BMI was observed within the full episode window.
21. `nadir_bmi` — Lowest EWMA-smoothed BMI observed within the full episode window.
22. `end_bmi_raw` — Raw BMI value at `end_day` from the EHR.
23. `end_bmi_smooth` — EWMA-smoothed BMI ($\alpha=0.2$) at `end_day`.
24. `weight_loss_max` — Maximum percent weight loss from episode start to nadir. This is the metric that determined episode qualification ($\geq 5\%$, or $\geq 2\%$ if `start_bmi_smooth` < 20). Validated: 0 of 30,938 episodes fall below their applicable detection threshold.
25. `death_event` — Binary flag indicating whether the patient died during follow-up (1 = died, 0 = censored).
26. `os_days` — Overall survival in days from anchor date to death date (if `death_event=1`), or to the later of the last recorded clinical contact date or last available BMI measurement (if censored). The later of the two dates is used for censored patients because the clinical last-contact field reflects a fixed system-pull date and can understate true follow-up for patients seen, or BMI-measured, after that pull.
27. `followup_days` — Identical to `os_days` by construction (see above).