

Temple (TUH) EEG: how many files, hours and patients

TOTAL on disk, nothing filtered, and below it how that narrows down to what reaches training. Hours are full recordings (no duration cap).

26,977

EDF files

6,703

hours of EEG

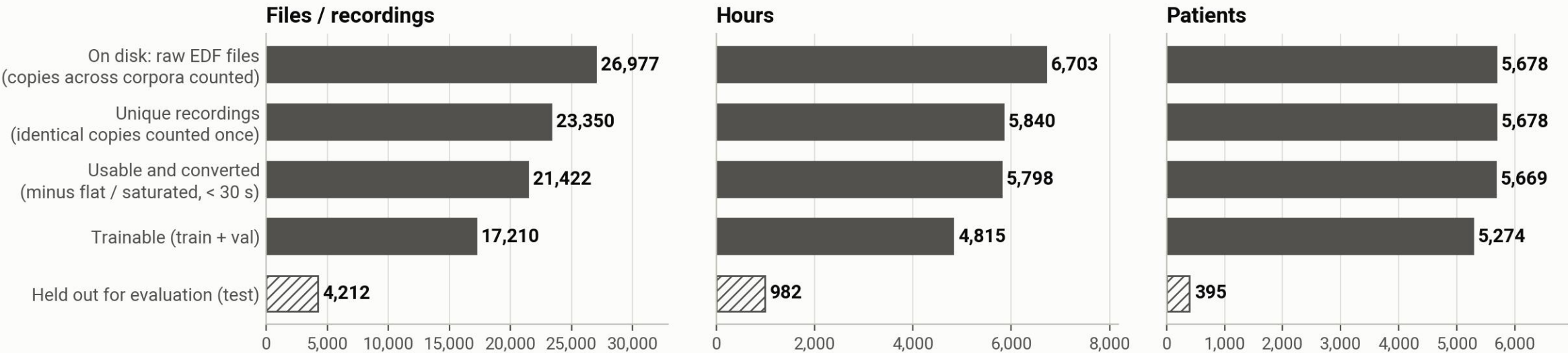
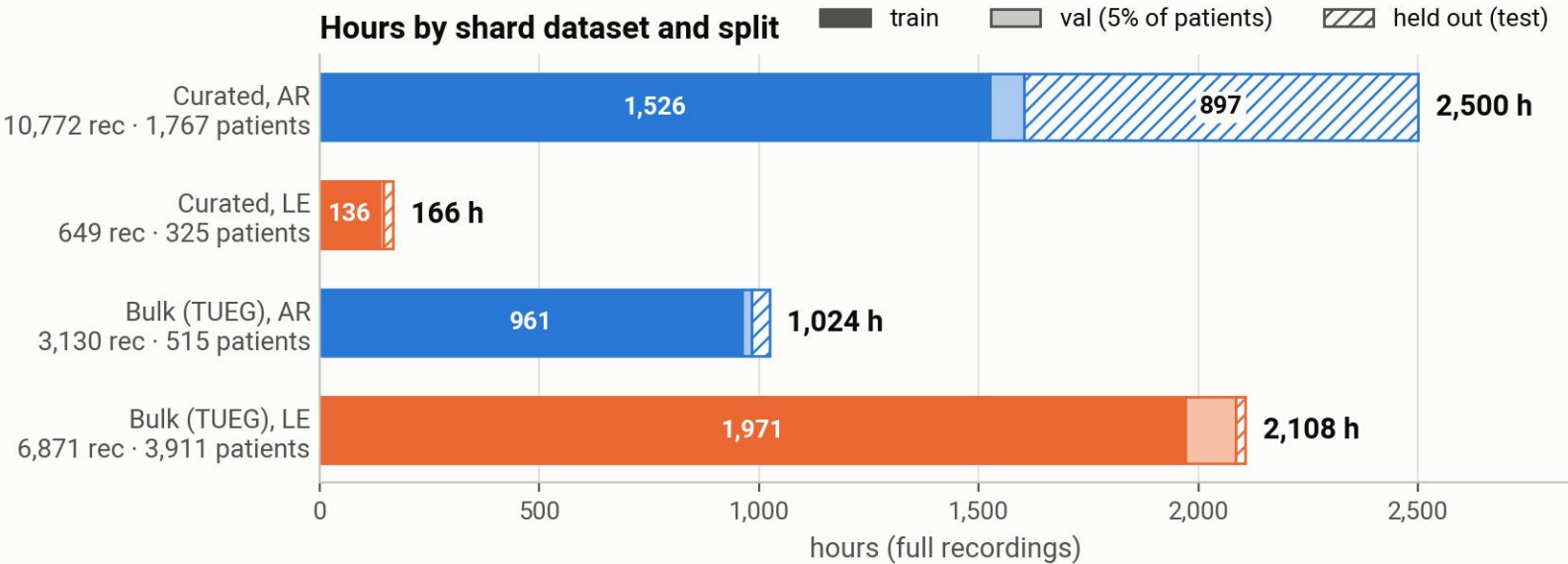
5,678

patients

420 GB

on disk

when the transfer finishes:
65,329 files · 23,209 h · 11,642 patients



How to read this

AR / LE = the hospital montage family:
AR is NOT an average reference (a single vertex-like reference); LE = linked ears.

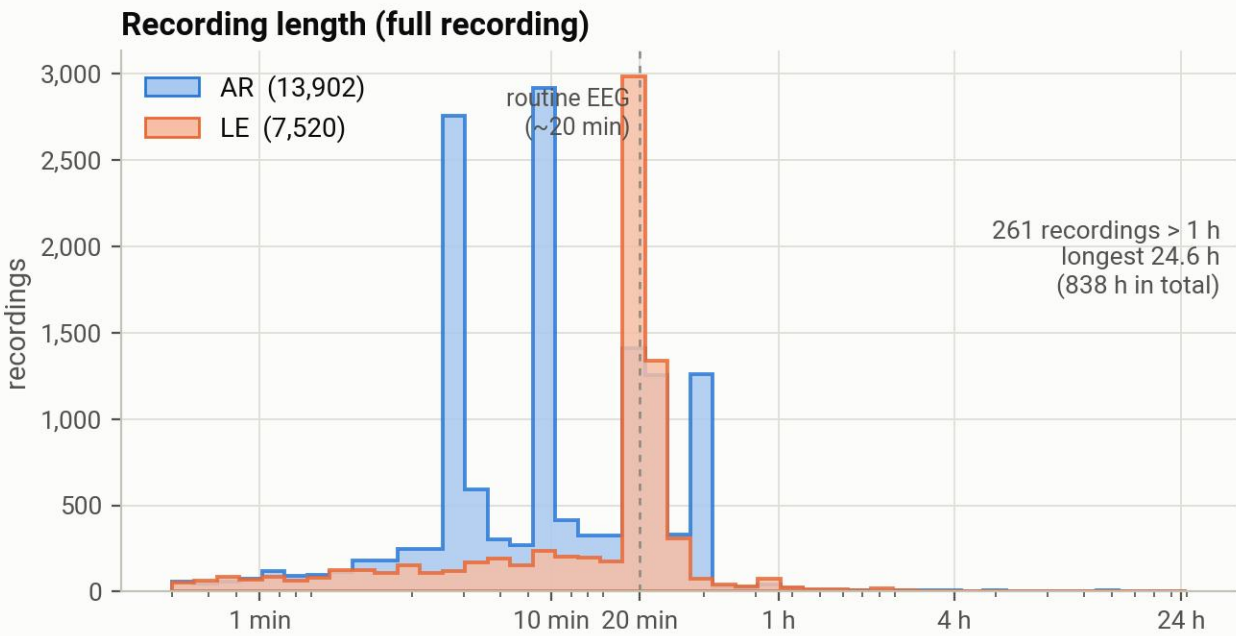
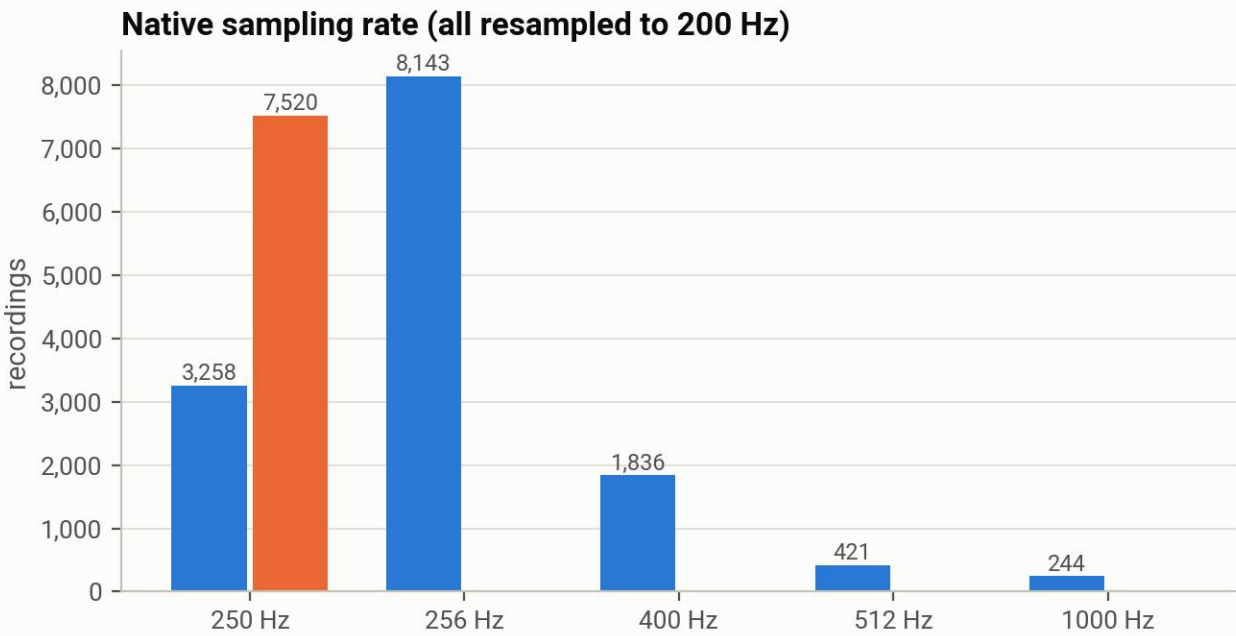
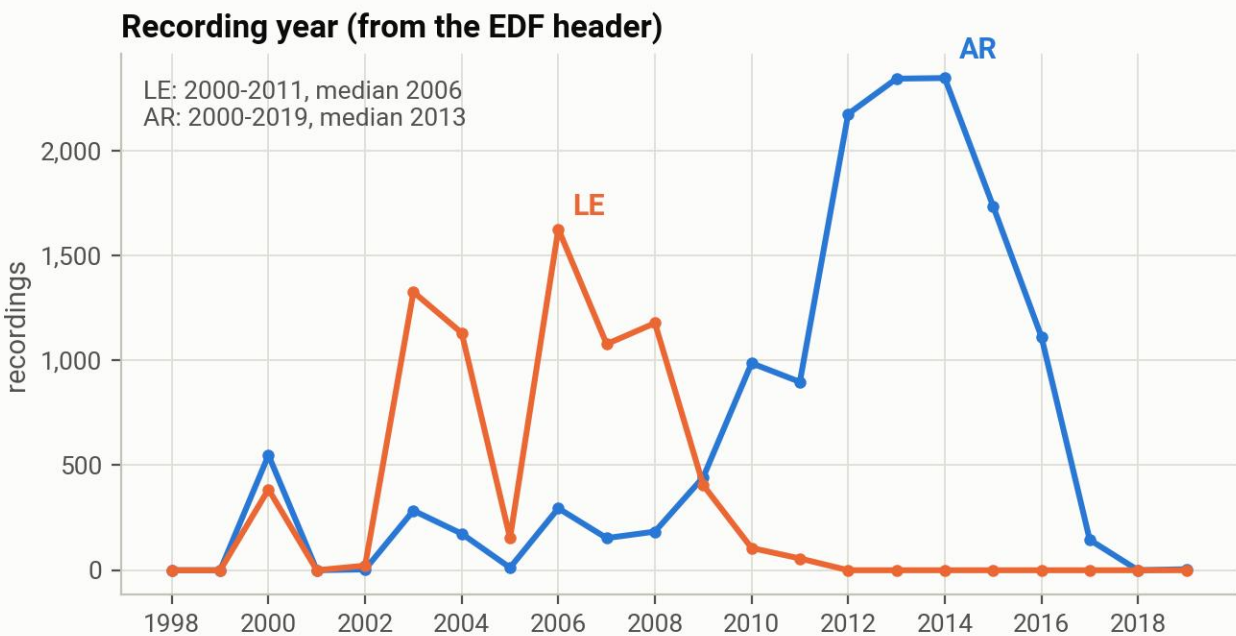
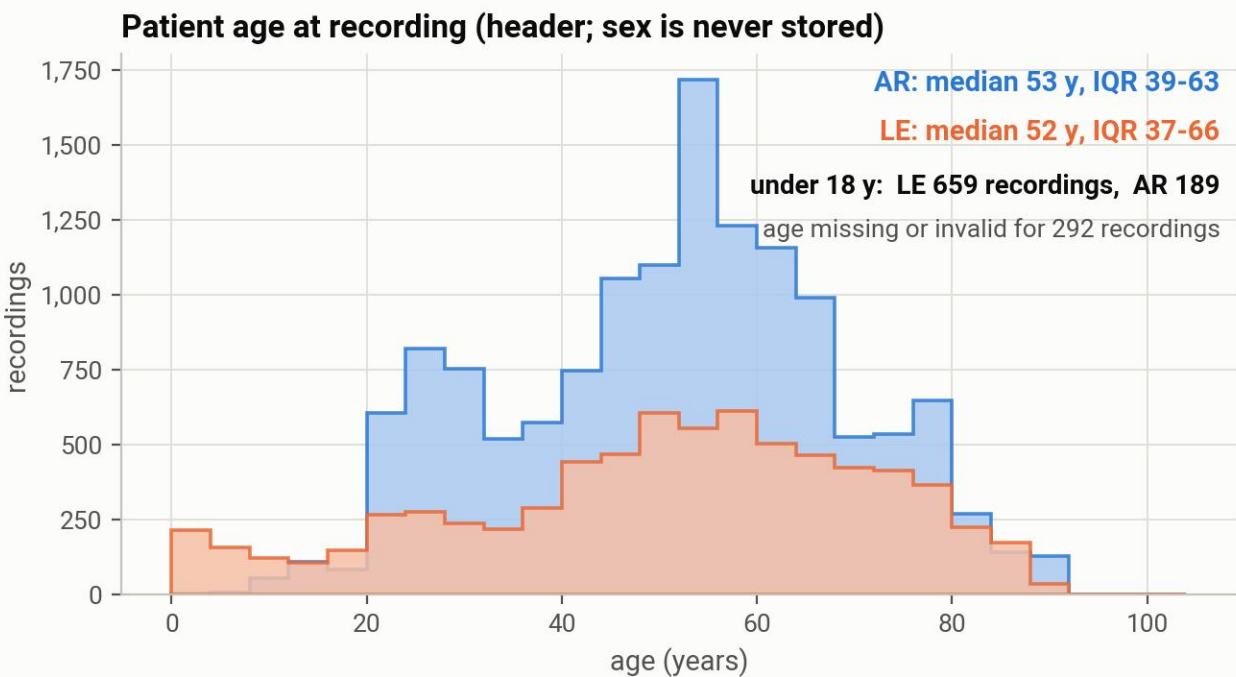
Held out = every recording of the 395 patients in TUAB eval, TUSZ eval/dev and TUEV eval (all corpora), so a patient never straddles train and test.

Not converted: 1,764 recordings < 30 s and 164 flat / saturated recordings.

Totals count a recording once per corpus it is in (TUSZ, TUEP, TUAB are subsets of TUEG). TUEV files carry no patient ids.

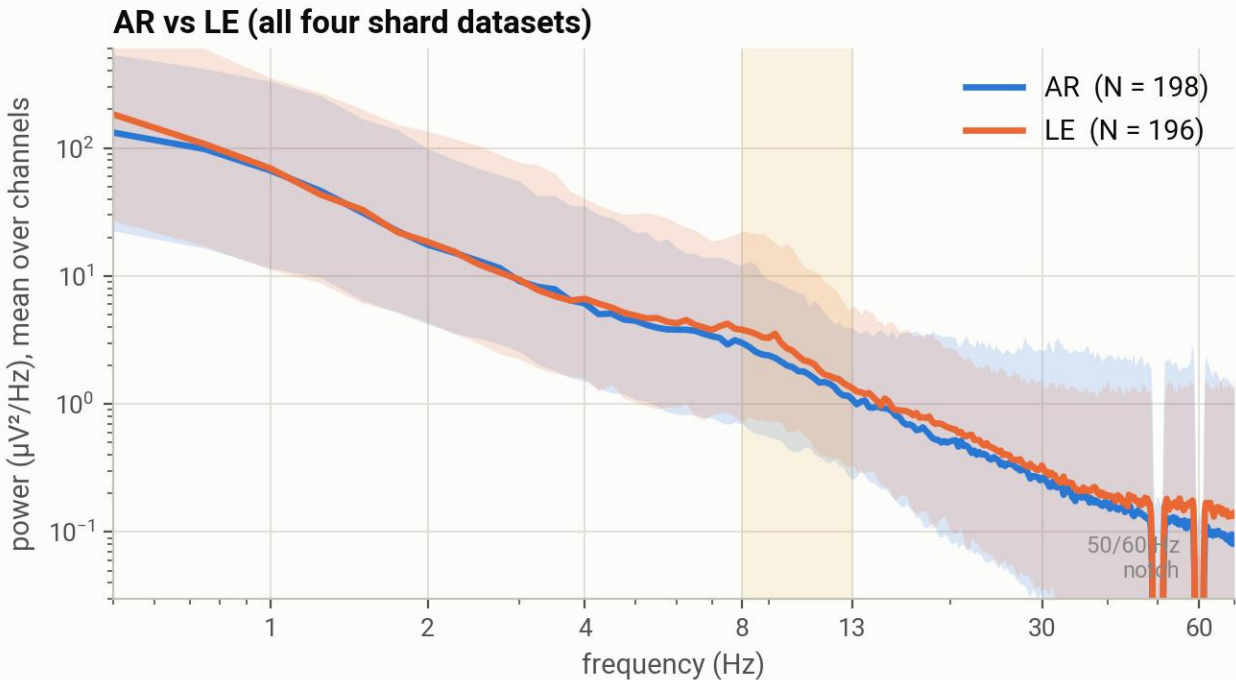
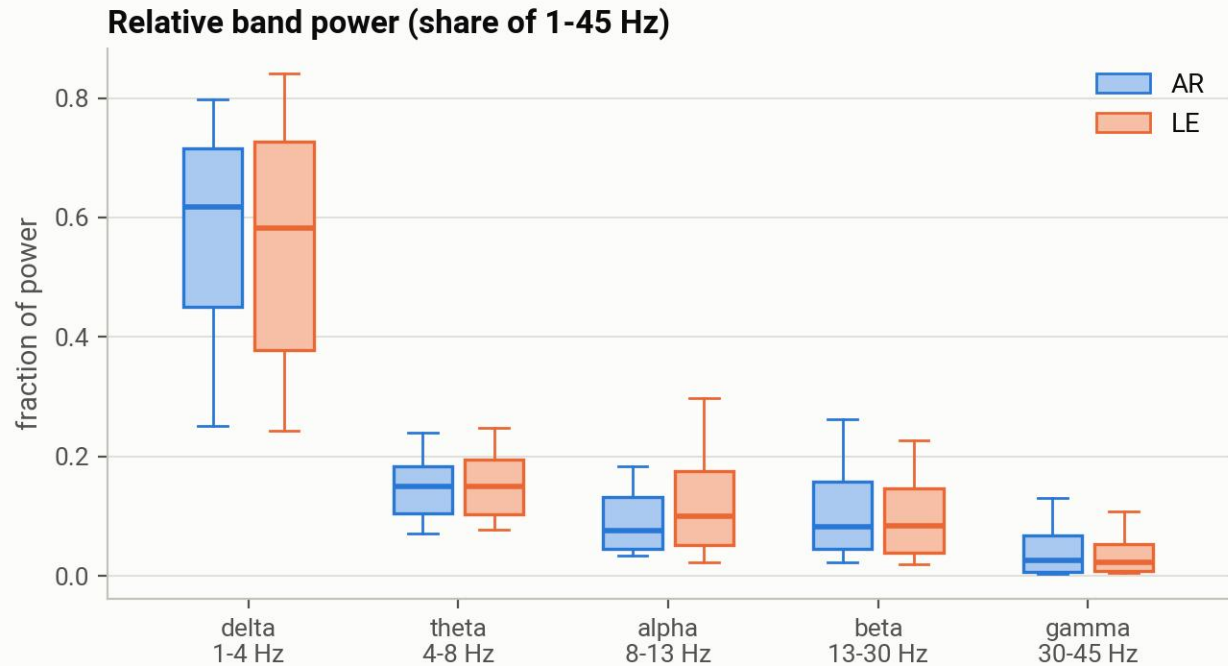
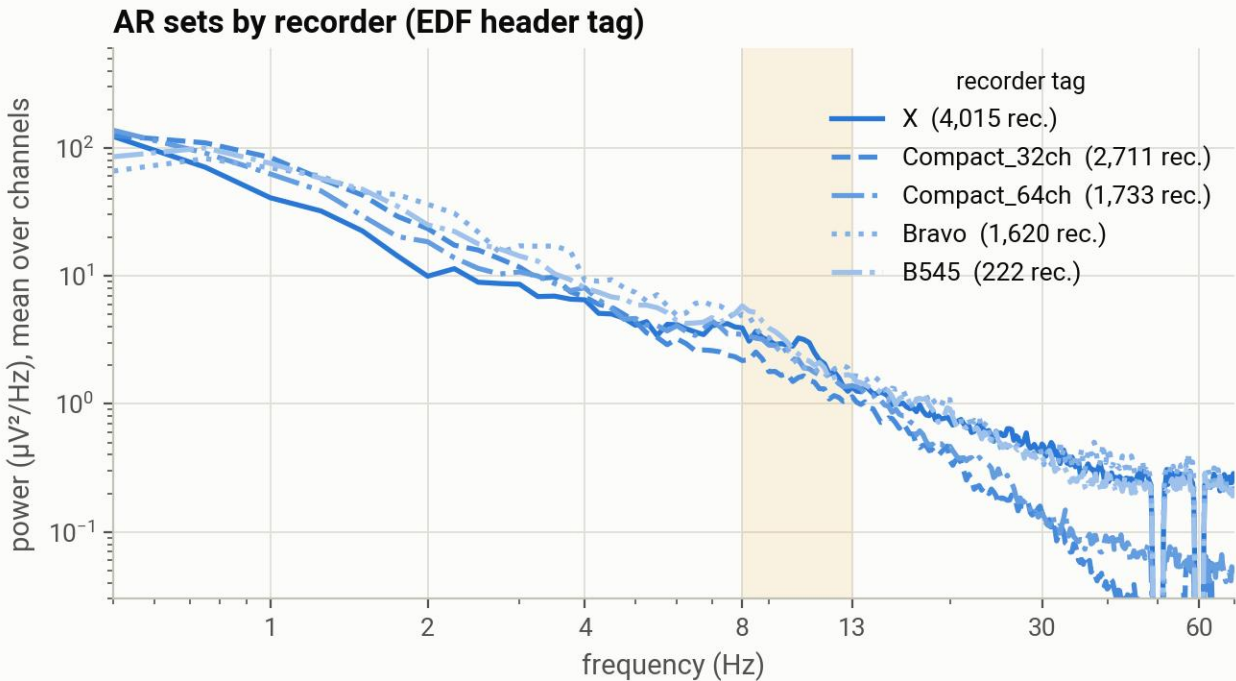
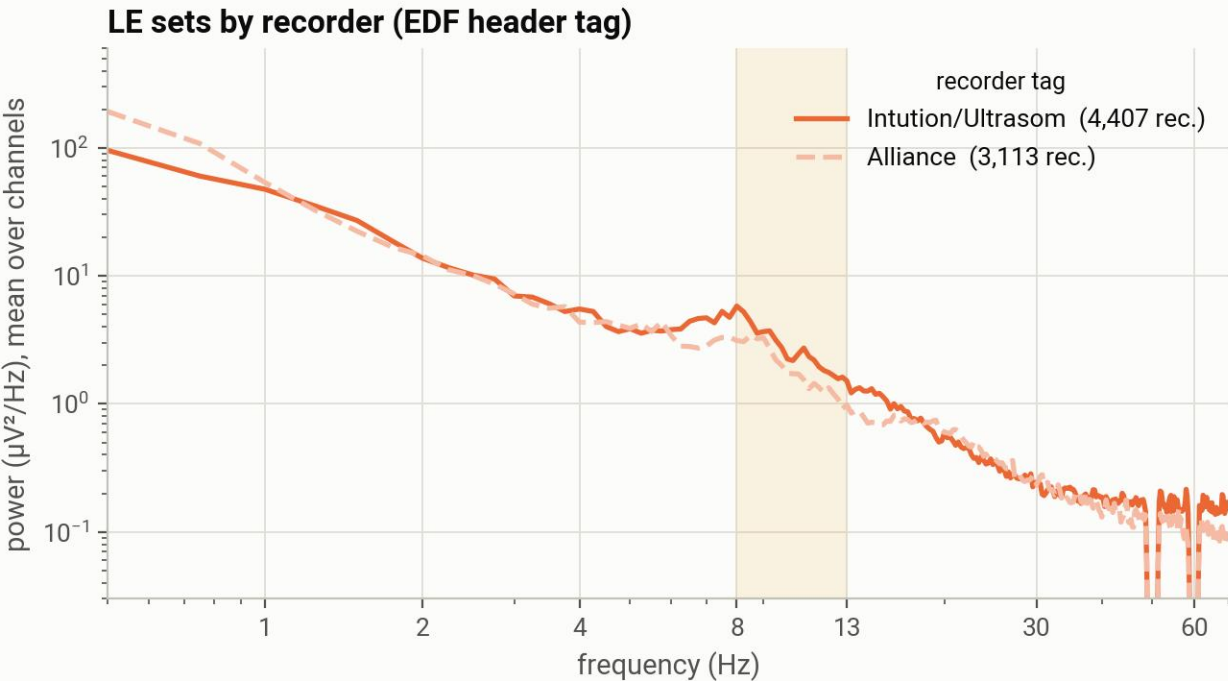
Temple recordings: length, sampling rate, era, age

AR and LE come from different recorders, eras and age groups (a confound to keep in mind). N = 21,422 converted recordings of the 26,977 files on disk, 5,669 patients



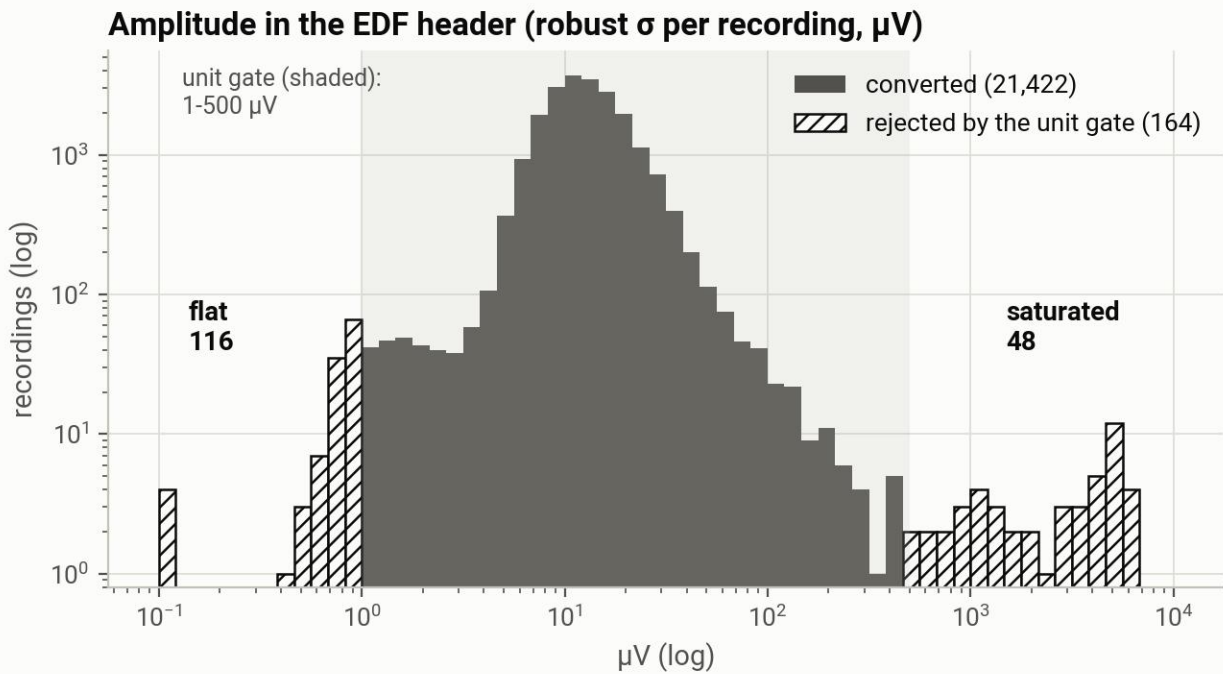
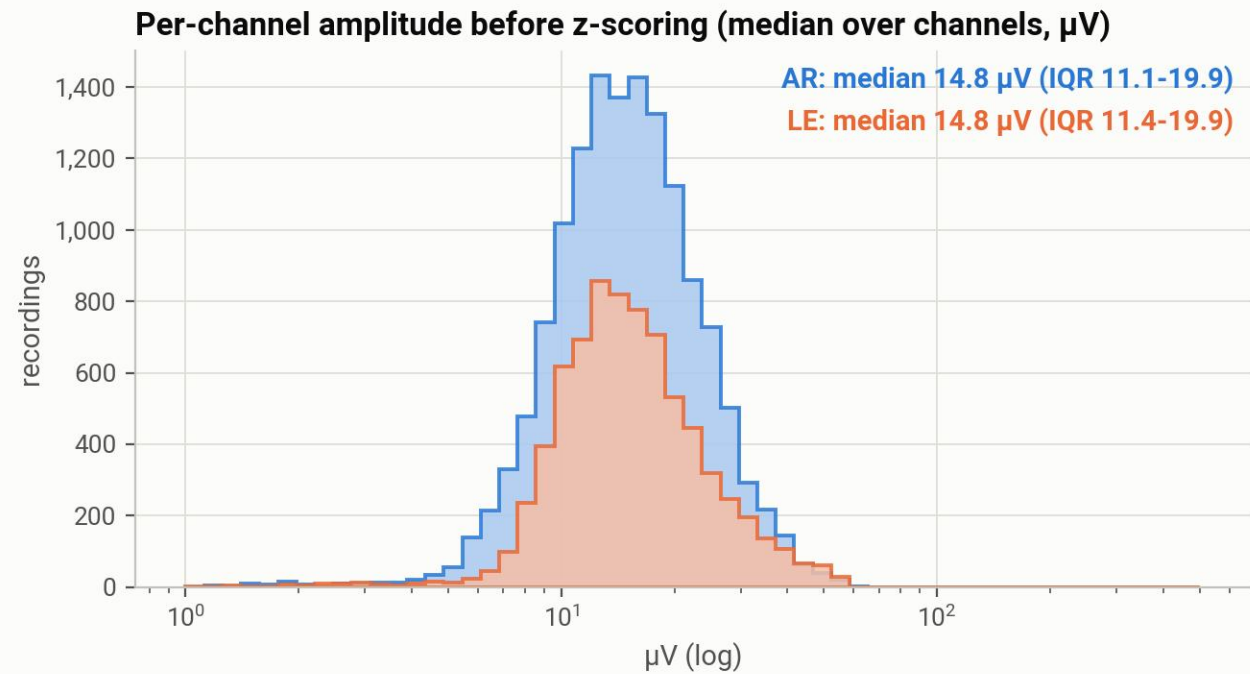
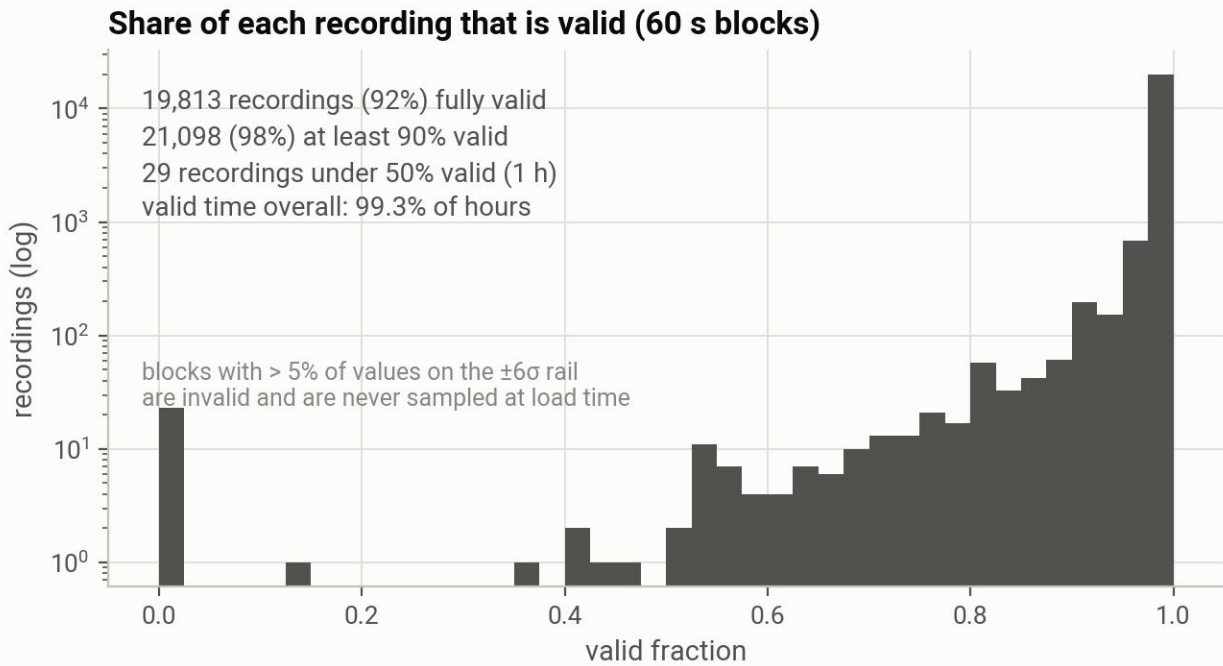
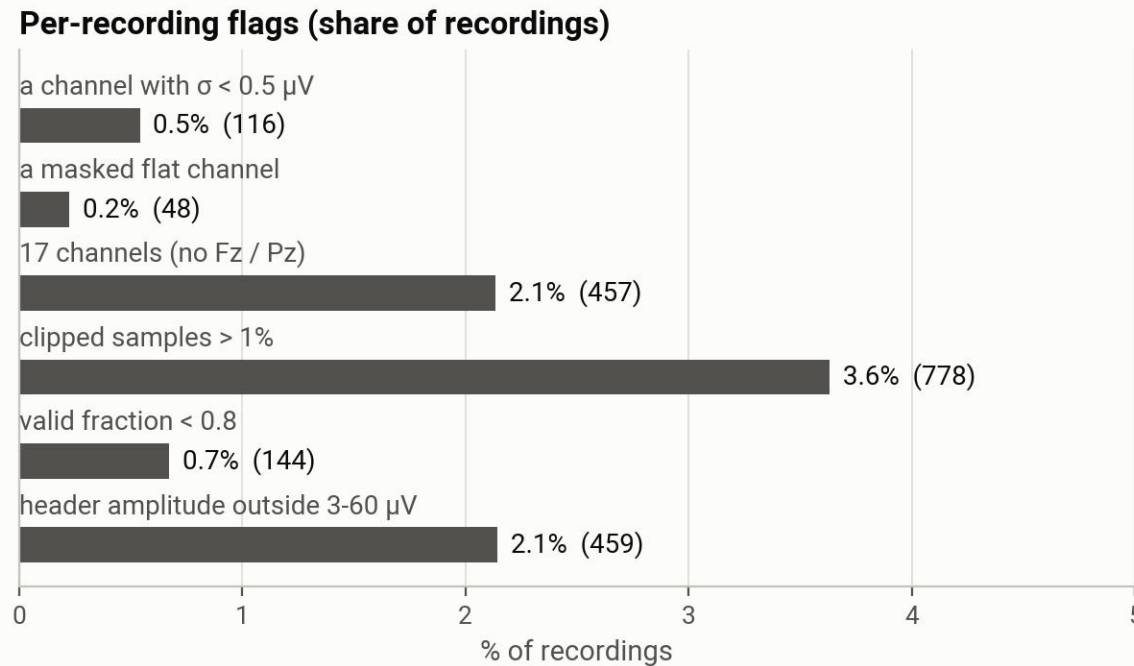
Temple signals: power spectra look like clinical EEG in every subset

Median over recordings with 10-90% band; one random 60 s valid window each (N = 394). AR and LE overlap; recorders differ above 30 Hz.



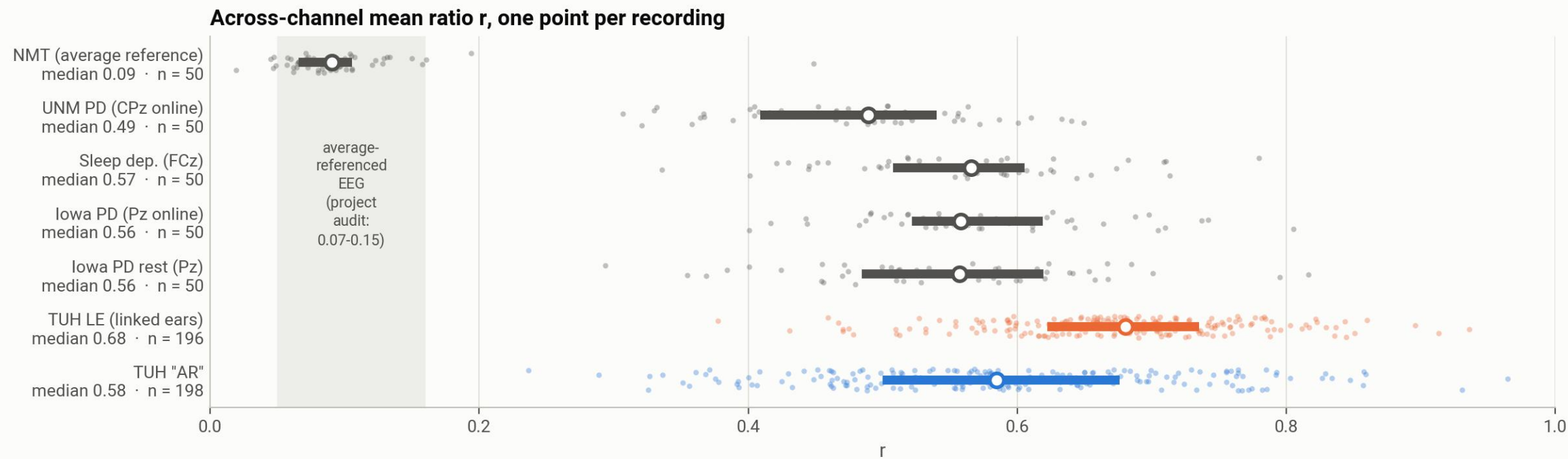
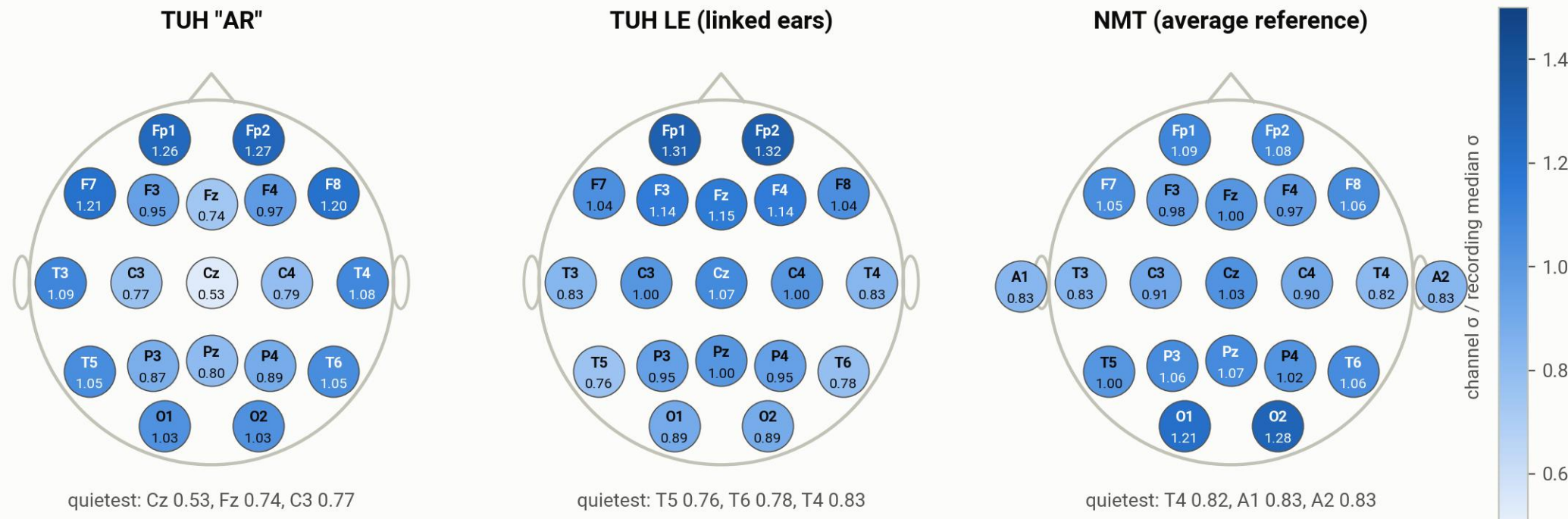
Temple signal checks: amplitude, units, valid time, channels

What the unit gate and the validity masks do. 26,977 files on disk; 21,422 recordings converted, 164 rejected by the unit gate



Temple "AR" is not an average reference

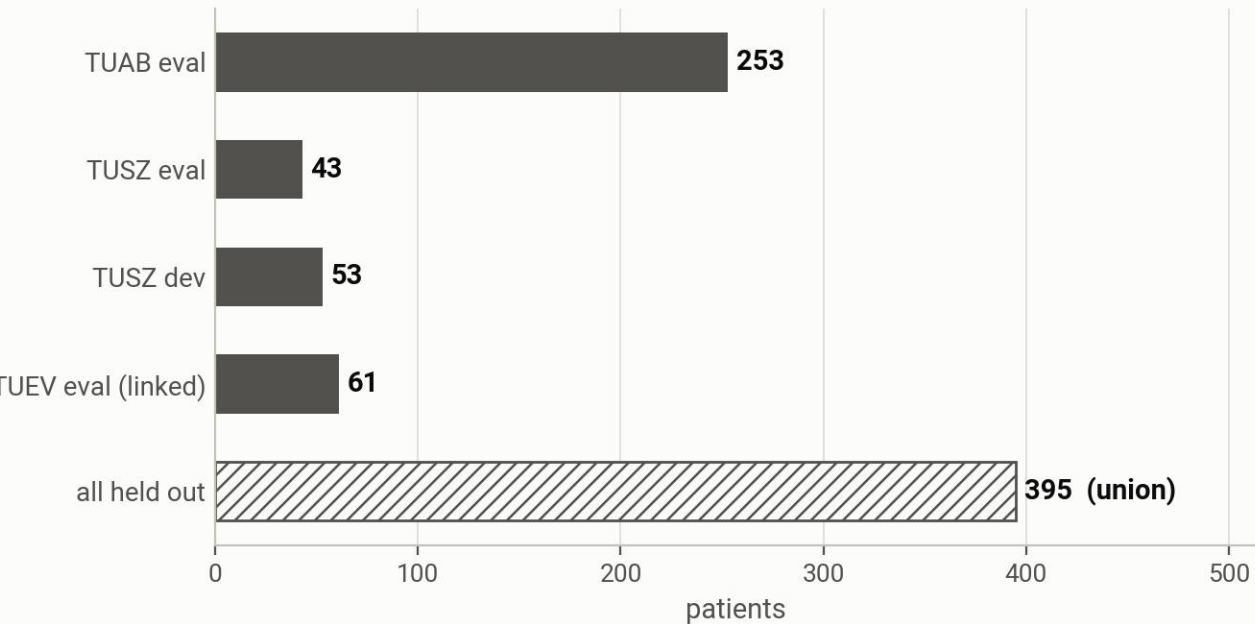
r = spread of the across-channel mean / mean channel spread (60 s windows): ~ 0.1 when the average was subtracted, higher with one physical reference. Per-channel amplitude maps below



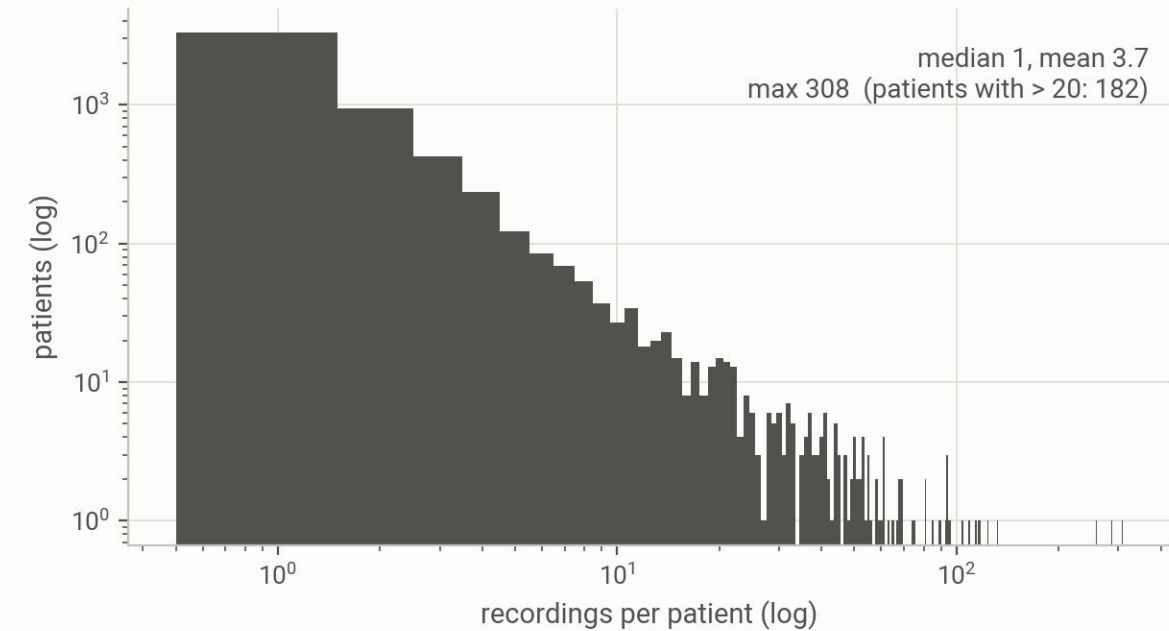
Temple patients: hold-out, recordings per patient, concentration

21,422 converted recordings from 5,669 patients (of 5,678 on disk); a patient never straddles train / val / test

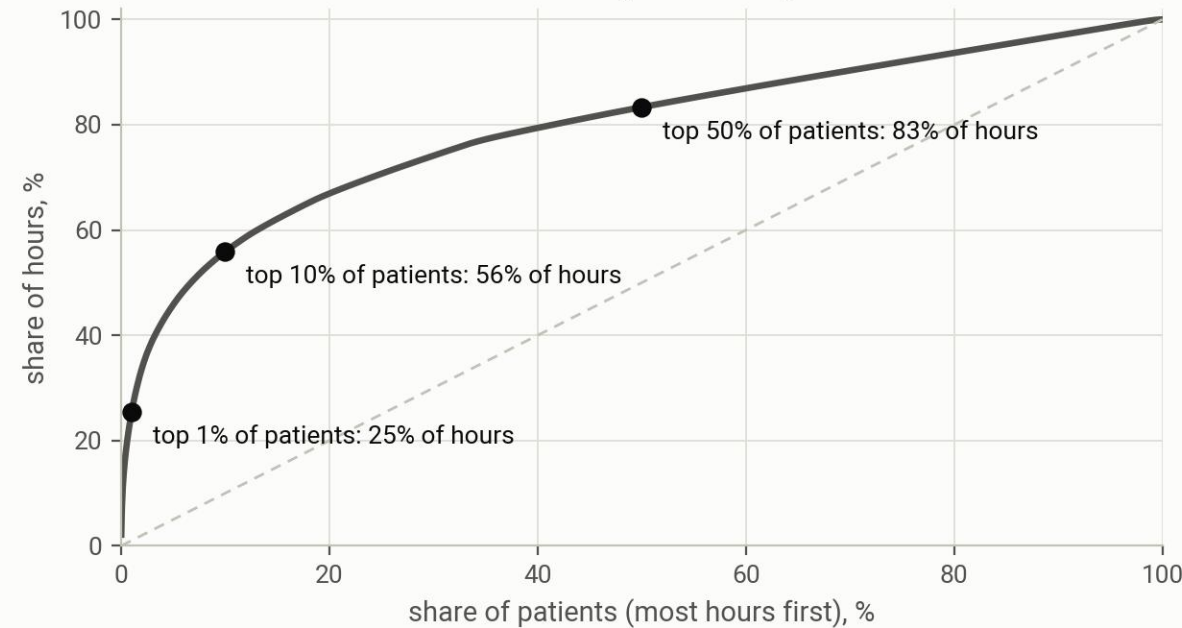
Held-out patients by evaluation set



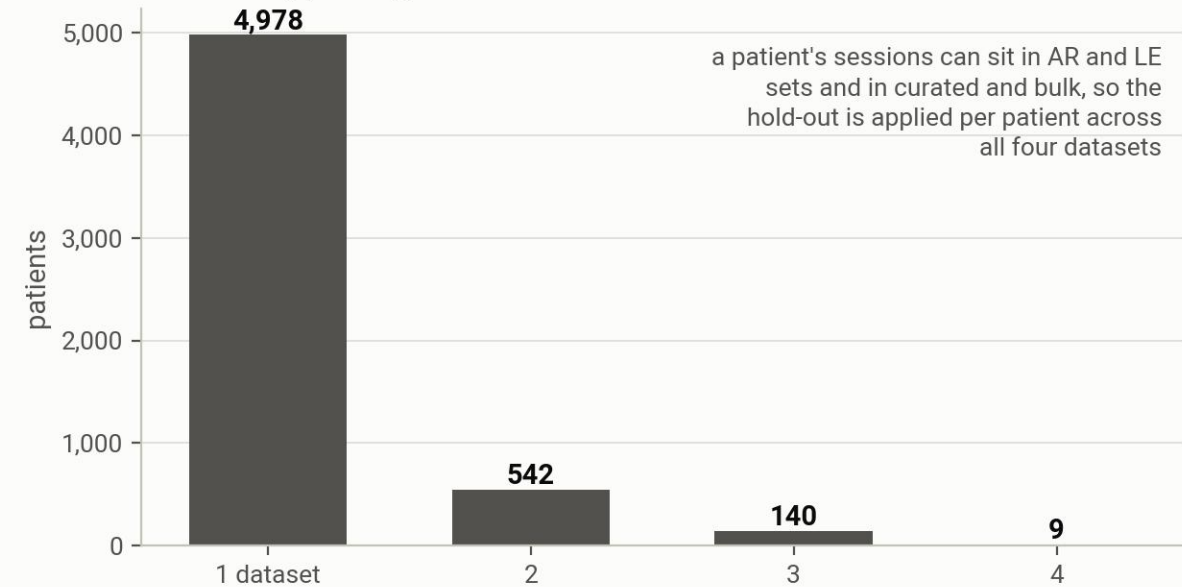
Recordings per patient



Concentration: how the hours are spread over patients



Patients appearing in several shard datasets

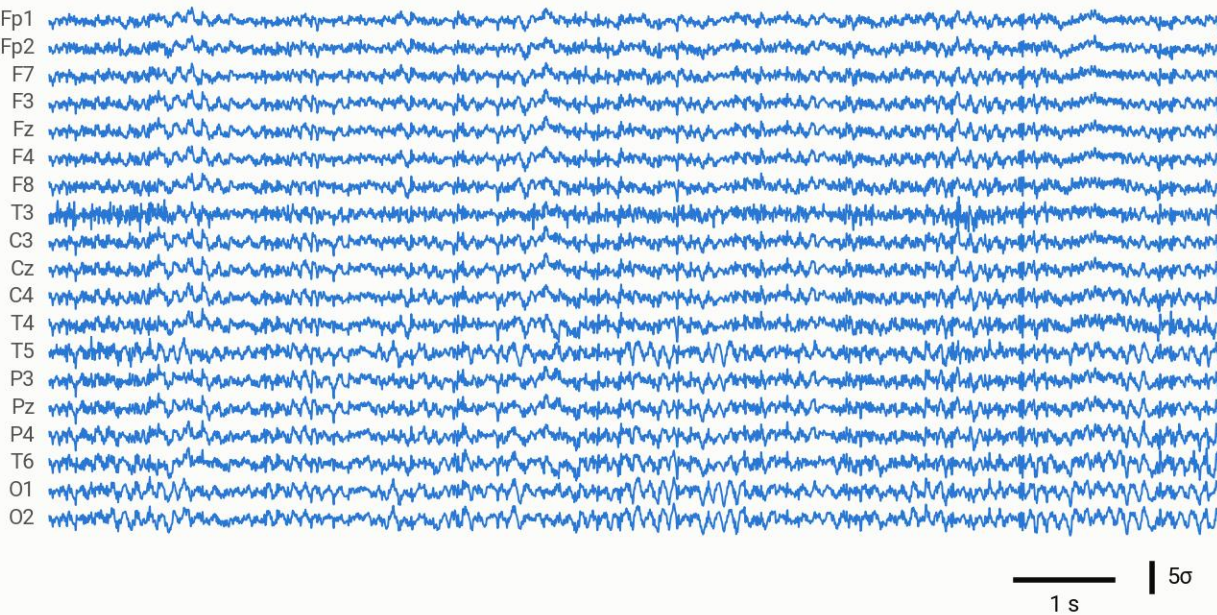


Temple example recordings (random draws, not selected)

One random recording per shard dataset, a 12 s window from the middle of its longest valid stretch; z-scored per channel; clipped at $\pm 6\sigma$ by construction

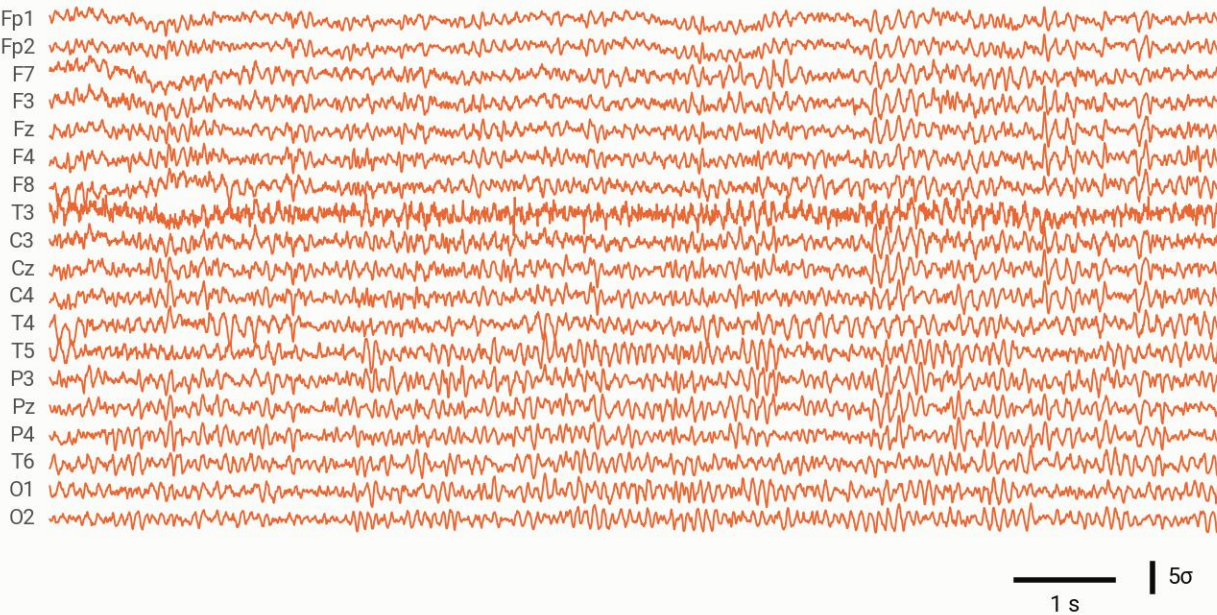
Curated, AR · recorder X · 250 Hz native · age 55

aaaaafge_s003_t000



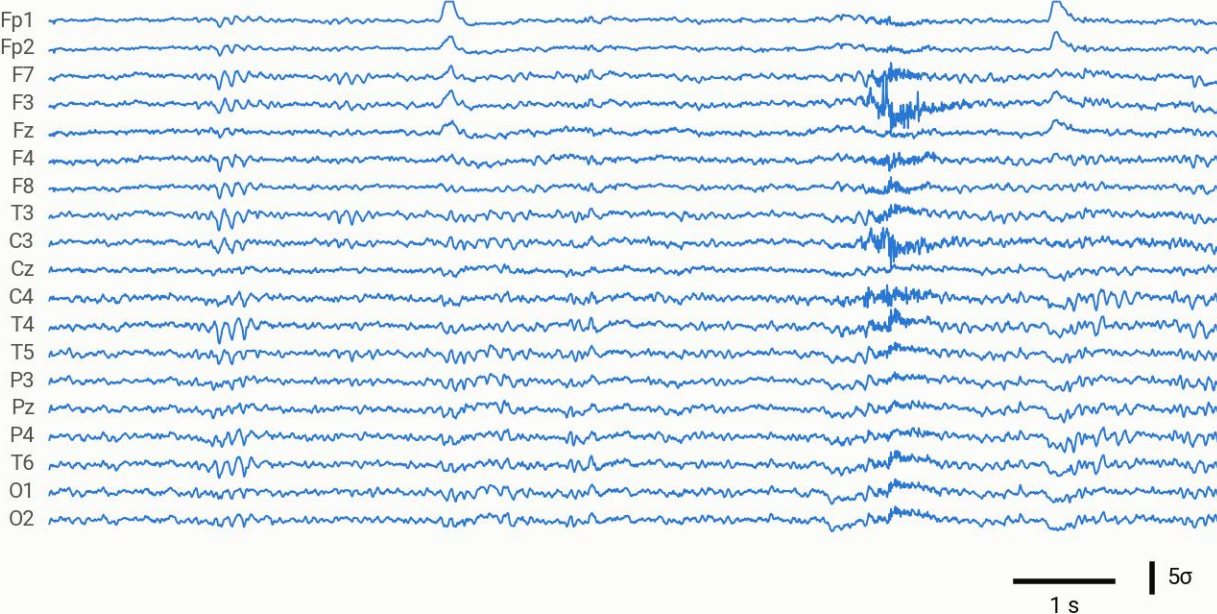
Curated, LE · recorder Intution/Ultrasom · 250 Hz native · age 58

aaaaaawu_s003_t000



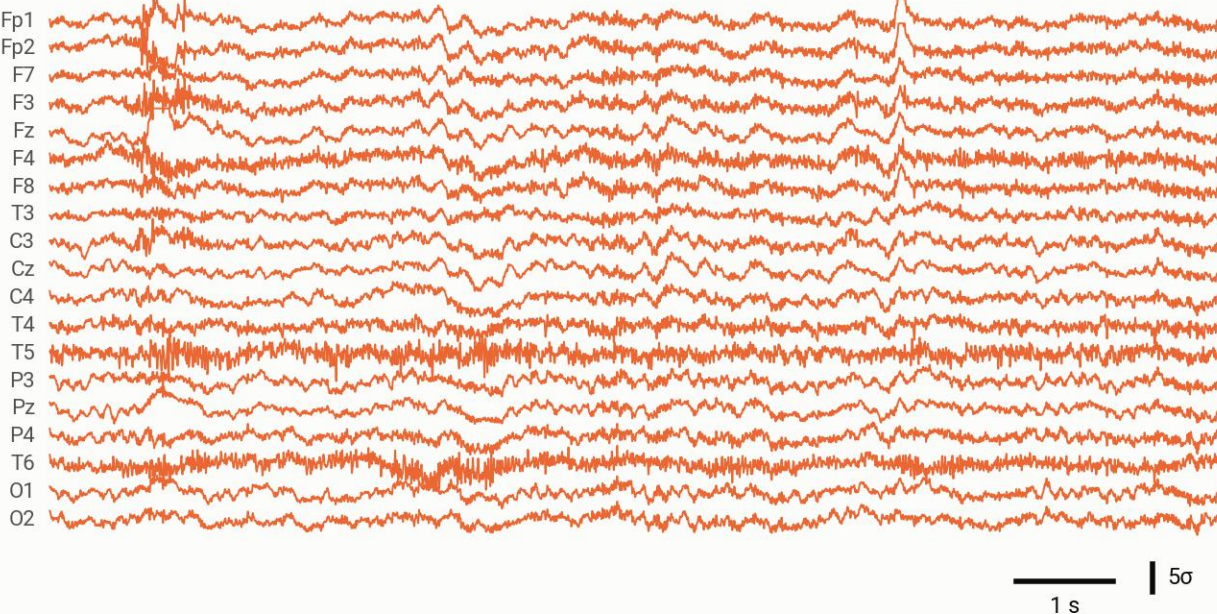
Bulk (TUEG), AR · recorder Compact_64ch · 256 Hz native · age 54

aaaaahsi_s007_t007



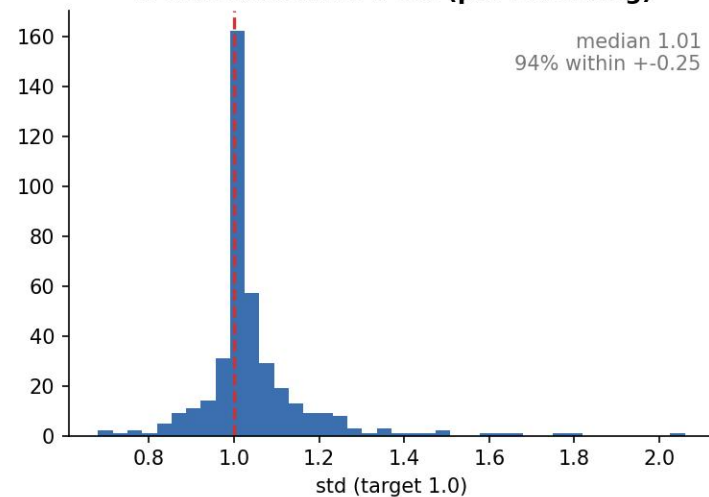
Bulk (TUEG), LE · recorder Alliance · 250 Hz native · age 2

aaaaadxy_s001_t003

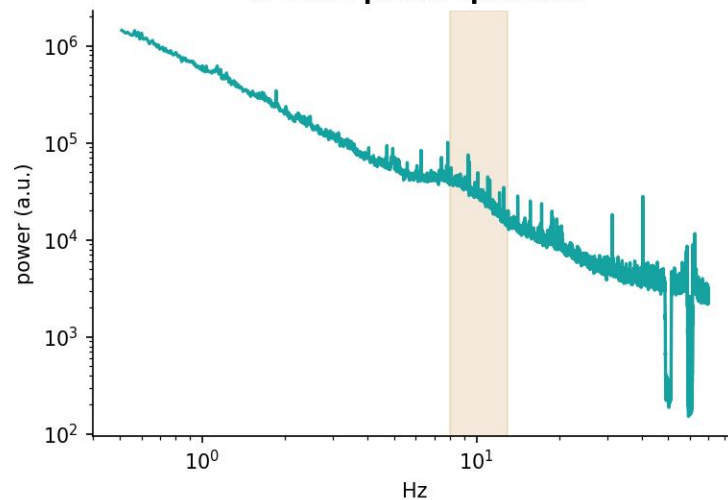


Quality control (shards) - TUH_CURATED_AR (10772 recordings, 1994 subjects, 2500 h)

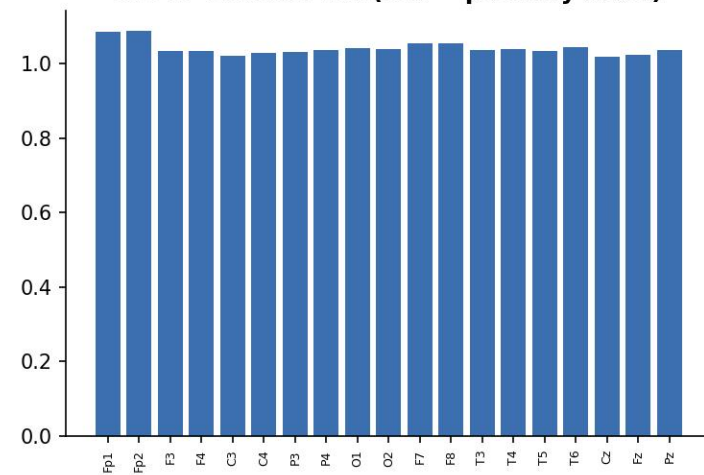
1. Normalization: z std (per recording)



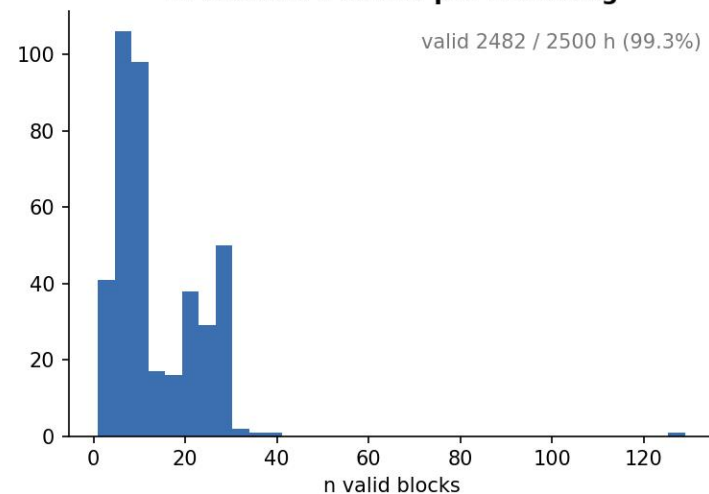
2. Mean power spectrum



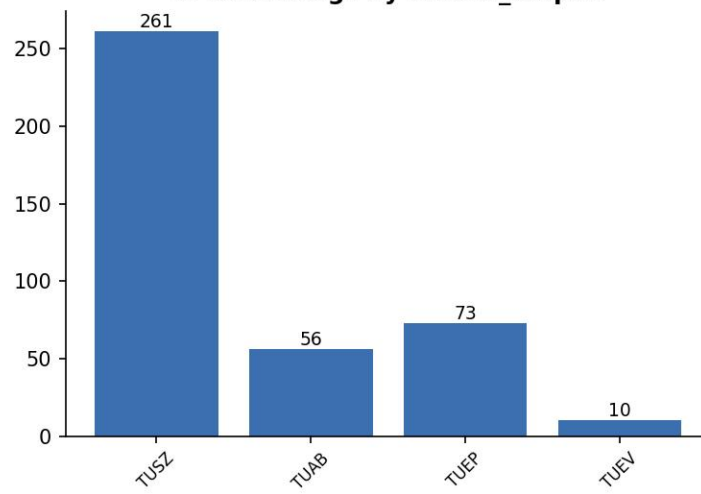
3. Per-channel std (red = possibly dead)



4. Valid 60 s blocks per recording



5. Recordings by source_corpus

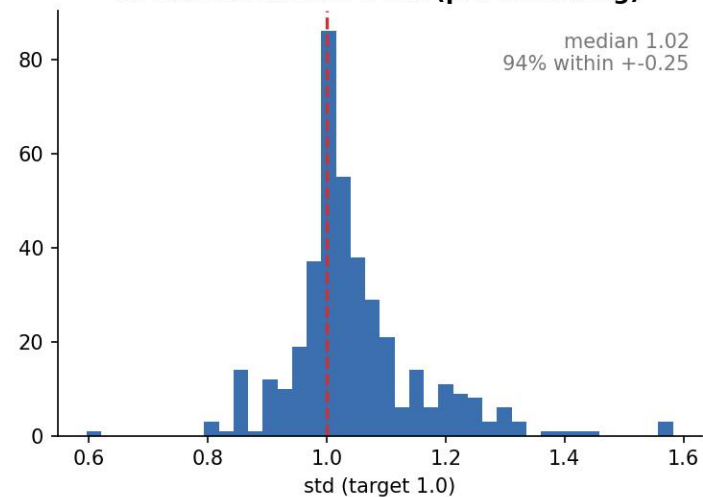


recordings inspected: 400
subjects (sample): 225
unreadable/NaN files: 0
mean clipped samples: 0.08%
dead channels: 0
unit status: resolved
exclude_from_mix: False
valid time: 99.3%

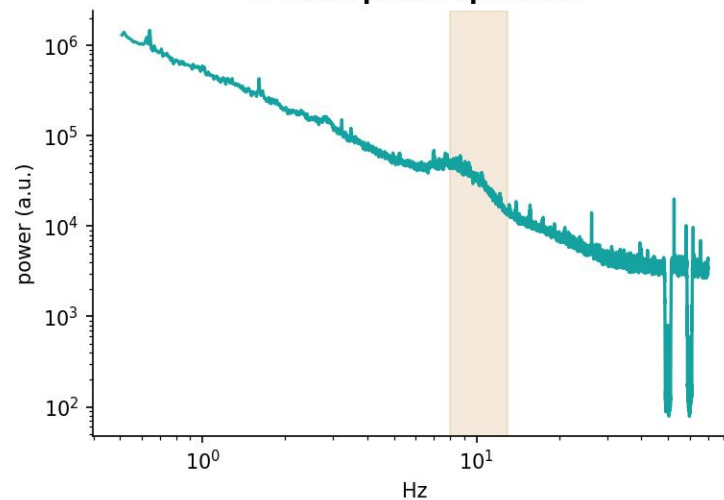
VERDICT:
PASS - safe to train on

Quality control (shards) - TUH_CURATED_LE (649 recordings, 325 subjects, 166 h)

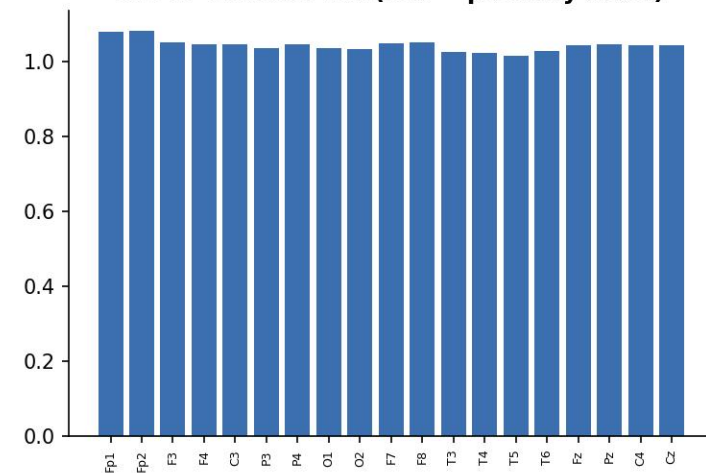
1. Normalization: z std (per recording)



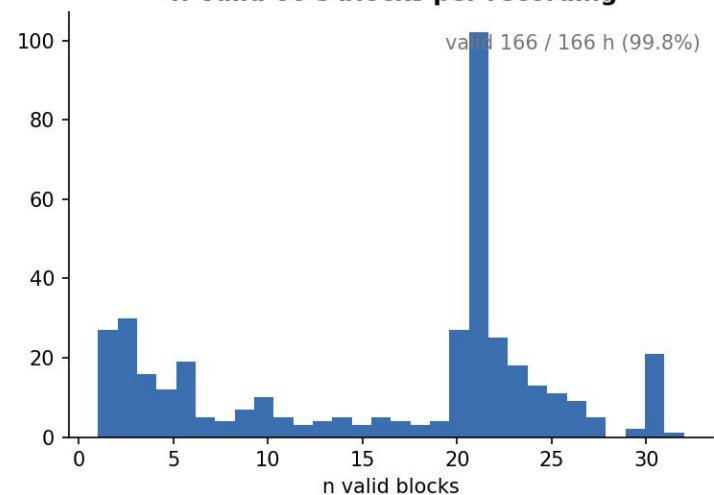
2. Mean power spectrum



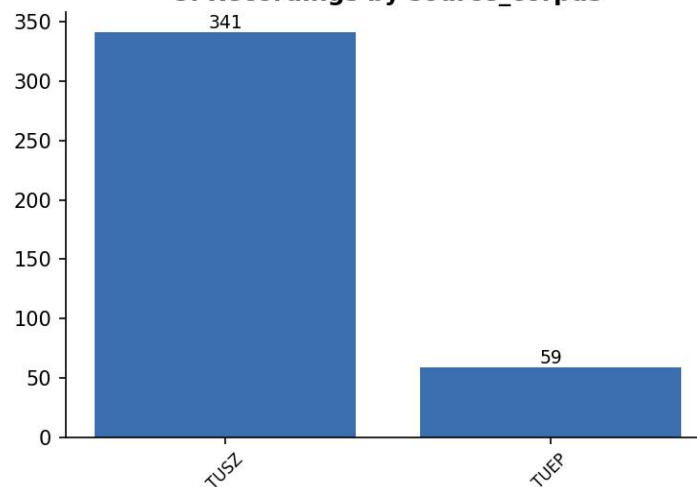
3. Per-channel std (red = possibly dead)



4. Valid 60 s blocks per recording



5. Recordings by source_corpus

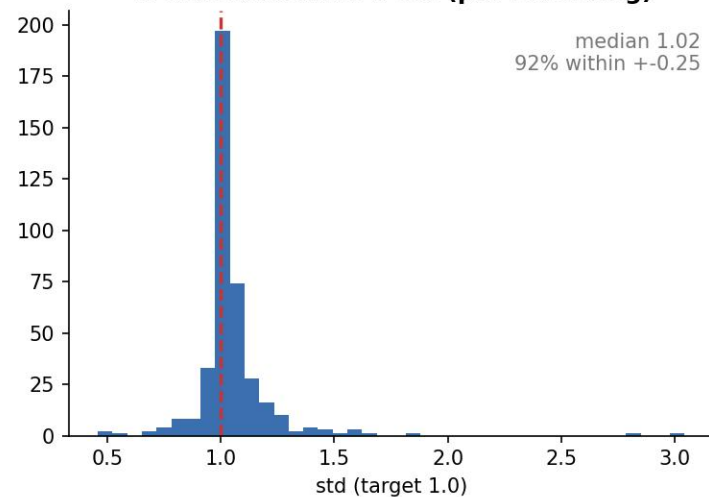


recordings inspected: 400
subjects (sample): 245
unreadable/NaN files: 0
mean clipped samples: 0.07%
dead channels: 0
unit status: resolved
exclude_from_mix: False
valid time: 99.8%

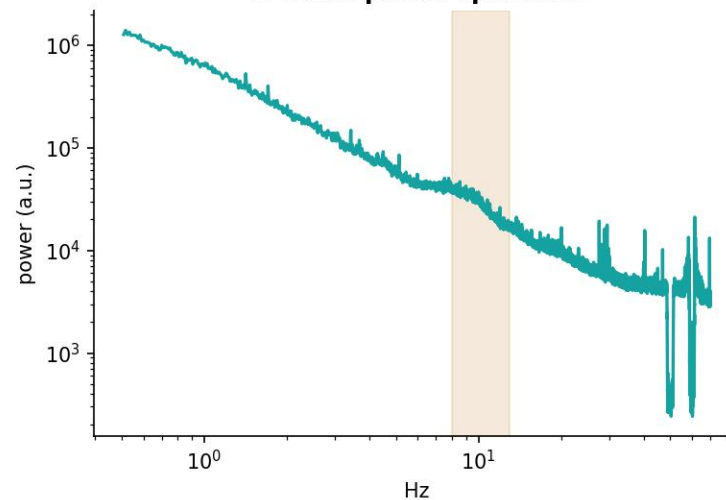
VERDICT:
PASS - safe to train on

Quality control (shards) - TUH_BULK_AR (3130 recordings, 515 subjects, 1024 h)

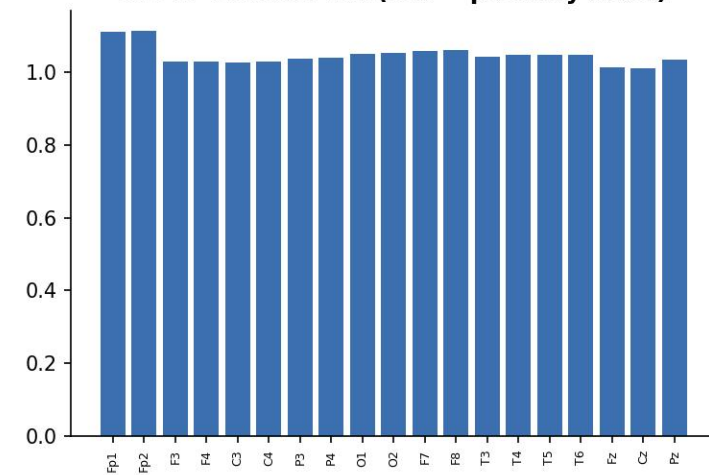
1. Normalization: z std (per recording)



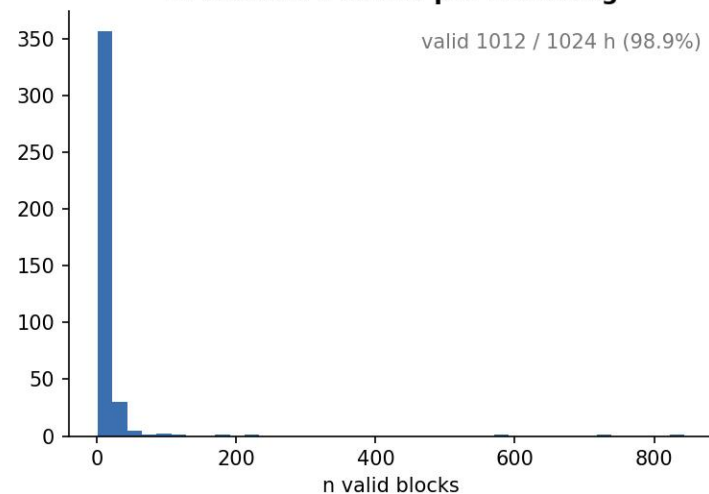
2. Mean power spectrum



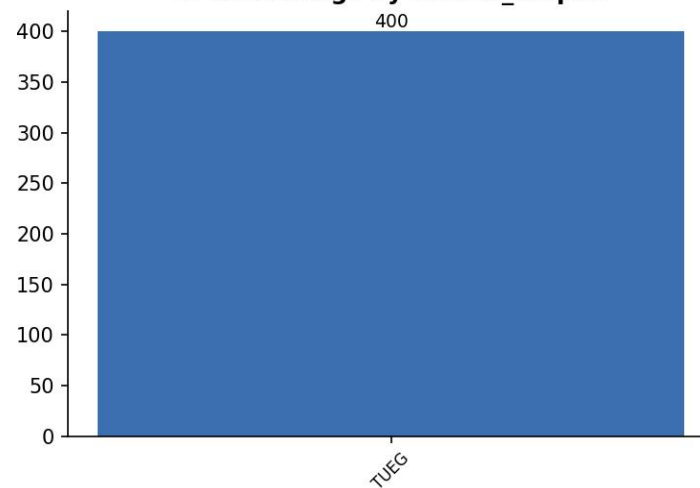
3. Per-channel std (red = possibly dead)



4. Valid 60 s blocks per recording



5. Recordings by source_corpus

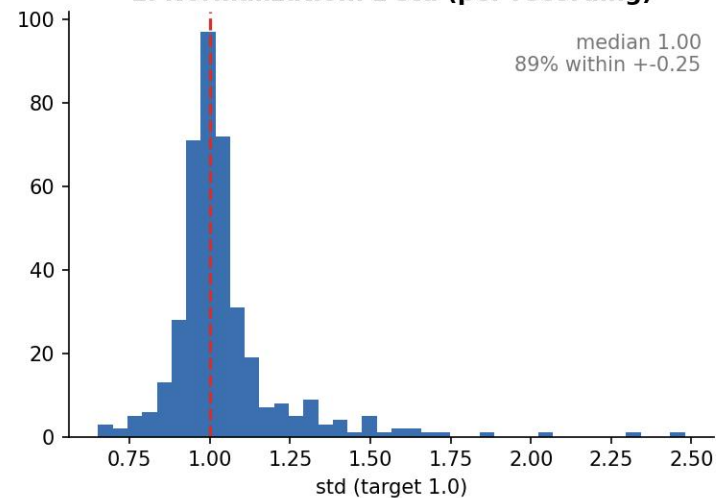


recordings inspected: 400
subjects (sample): 192
unreadable/NaN files: 0
mean clipped samples: 0.13%
dead channels: 0
unit status: resolved
exclude_from_mix: False
valid time: 98.9%

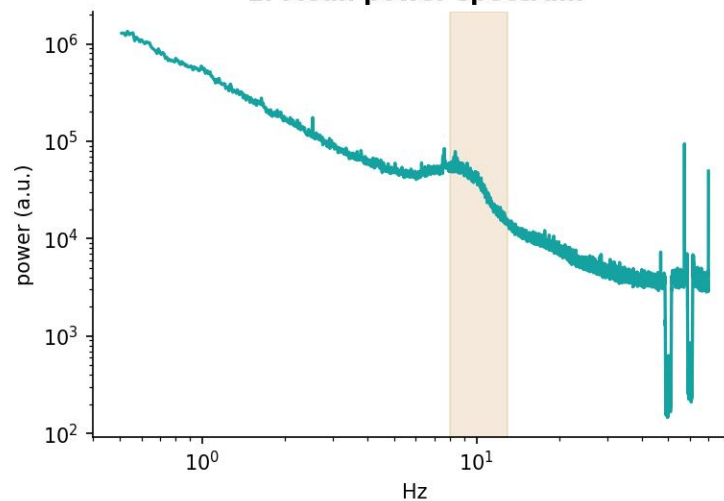
VERDICT:
PASS - safe to train on

Quality control (shards) - TUH_BULK_LE (6871 recordings, 3911 subjects, 2108 h)

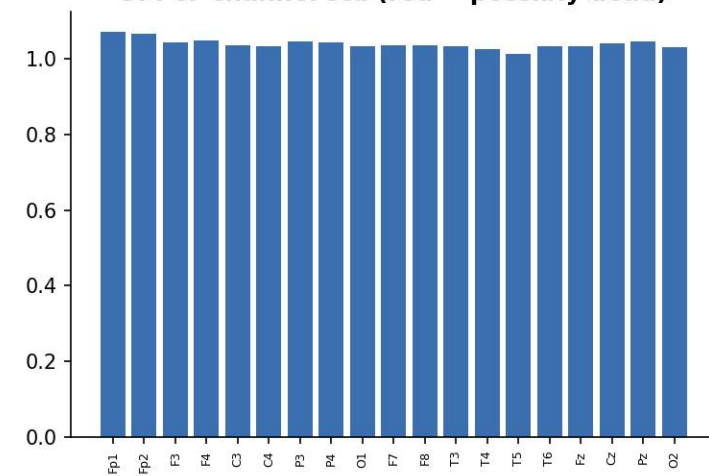
1. Normalization: z std (per recording)



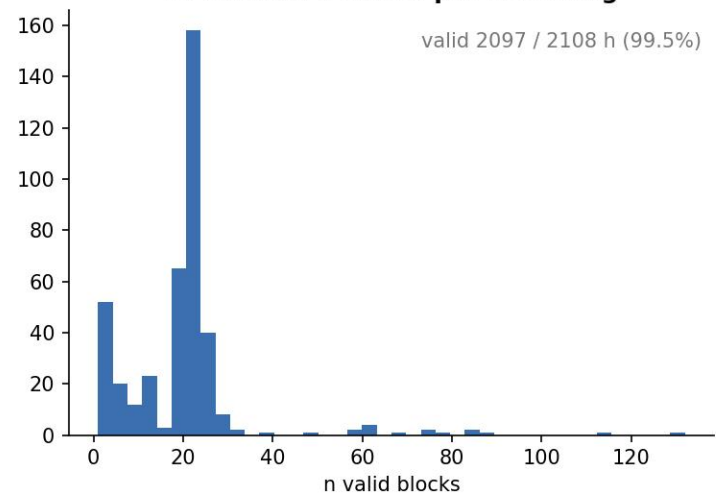
2. Mean power spectrum



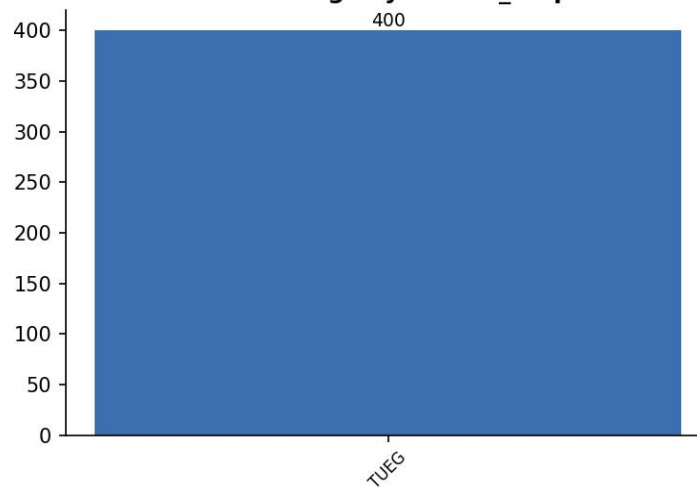
3. Per-channel std (red = possibly dead)



4. Valid 60 s blocks per recording



5. Recordings by source_corpus



recordings inspected: 400
subjects (sample): 380
unreadable/NaN files: 0
mean clipped samples: 0.10%
dead channels: 0
unit status: resolved
exclude_from_mix: False
valid time: 99.5%

VERDICT:
PASS - safe to train on