On the choice and influence of the number of boosting steps

Heidi Seibold, Christoph Bernau, Anne-Laure Boulesteix and Riccardo De Bin

Epidemiology, Biostatistics and Prevention Institute (EBPI)
University of Zurich

Department of Medical Informatics, Biometry and Epidemiology (IBE)
University of Munich (LMU)

Boosting

- Interpretable results
- Good prediction models
- Can handle high dimensional data (n $<<$ p)
  $\rightarrow$ shrinkage and variable selection
- Handles several types of outcomes $\Rightarrow$ Here: right-censored time to event data under proportional hazards assumption

User friendly R-software for time to event data:

- mboost (Hothorn et al., 2006)
- CoxBoost (Binder and Schumacher, 2008)

Boosting: Idea

Find a good model by iteratively updating

$$F_m(x, \theta) = F_{m-1}(x, \theta) + \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i, \theta) + f(x_i, \theta)\right)$$

θ is the parameter vector we want to optimize
f is a base learner
L is a loss function

Boosting for right-censored time to event data

	mboost	CoxBoost
loss function	negative partial log-likelihood	negative penalized partial log-likelihood (L_2 penalty)
base learner	least square estimator	first order approximation of the ML-estimator

Parameter tuning

Important tuning parameter:

Number of boosting steps m_{stop}

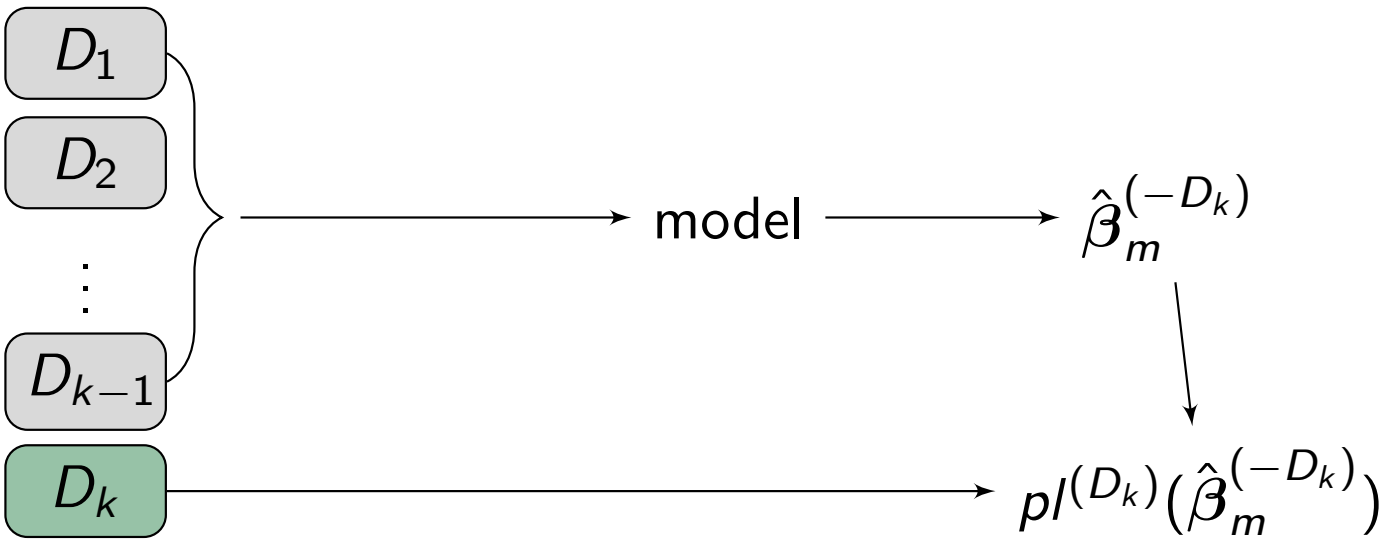
$\rightarrow$ Controls complexity

$\rightarrow$ Influences prediction accuracy

chosen often through cross-validation.

cross-validation for parameter tuning

k-fold CV; $m = 1, ..., m_{max}$



$$\Rightarrow \; cvpl(m) = \sum_{k=1}^{K} pl^{(D_k)}(\hat{\boldsymbol{\beta}}_m^{(-D_k)})$$

$$\Rightarrow \; \text{Choose } m_{stop} = \arg\max_m cvpl(m)$$

Note: *cvpl* is defined slightly different for CoxBoost. Idea is the same.

Cross-validation for parameter tuning: Problem

Cross-validation is random
(except leave-one-out $K = n$)

This influences:

- $\rightarrow$ Choice of m_{stop}
- $\rightarrow$ Number of chosen predictors
- $\rightarrow$ Prediction accuracy

Question: How much influence?

Empirical study

Conducted using 4 data sets
In each data set, for both mboost and CoxBoost:

- Conduct 3-, 5-, 10- and 20-fold CV; each 2000 times
- Conduct leave-one-out (loo) CV

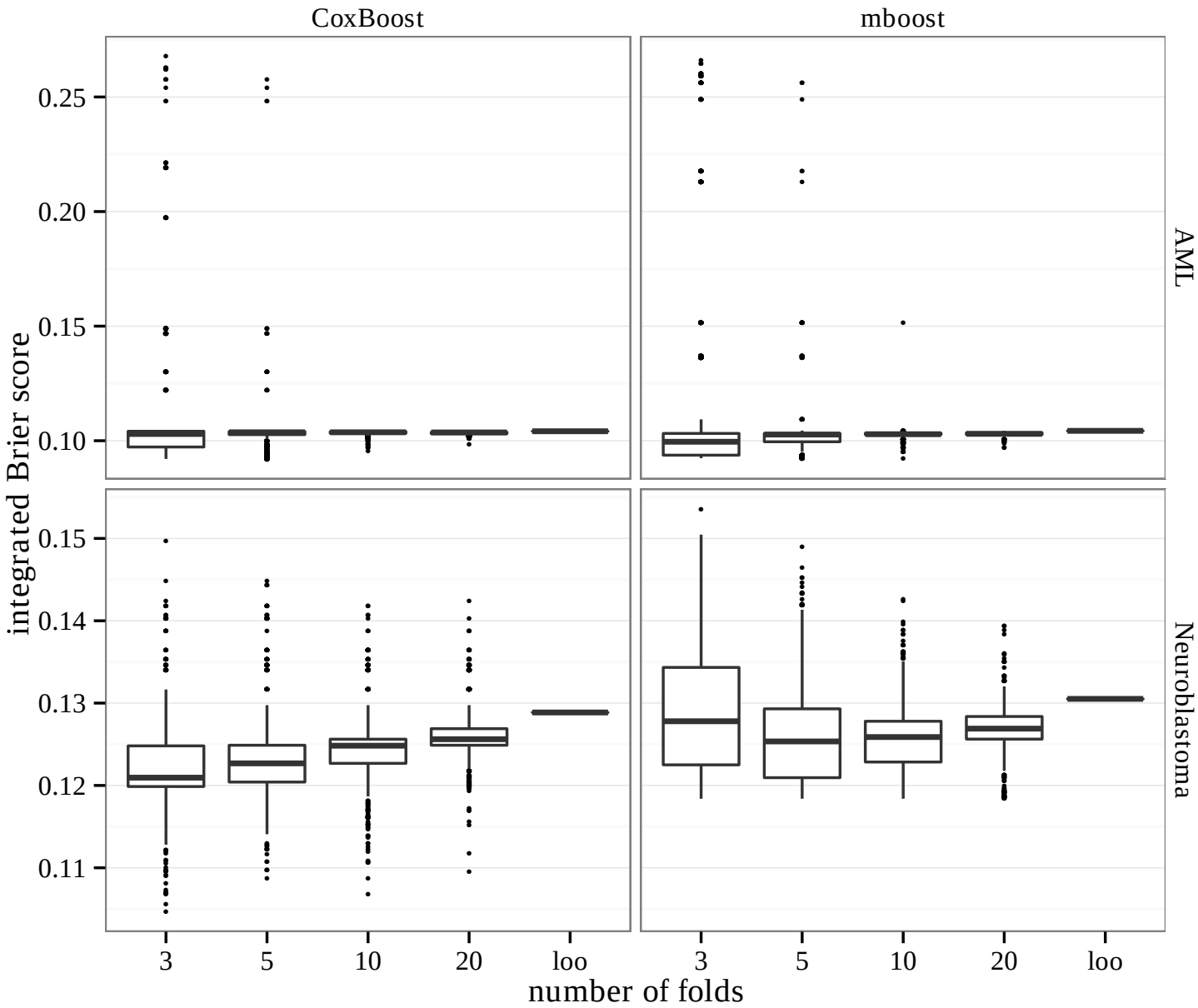
$\rightarrow$ Extract selected predictors

$\rightarrow$ Evaluate prediction accuracy on test set via (integrated) Brier score

Datasets

	reference	observations (uncensored) training set	validation set	predictors
Metzeler et al. (2008)		163 (103)	79 (32)	44754
Oberthuer et al. (2008)		242 (40)	120 (35)	9978

- Publicly available data
- Presence of independent training and validation sets
- Time-to-event response
- Gene expression data
- 2 different diseases (acute myeloid leukemia (AML), neuroblastoma)


CoxBoost
mboost
AML
Neuroblastoma
integrated Brier score
0.25
0.20
0.15
0.10
0.15
0.14
0.13
0.12
0.11
3
5
10
20
loo
number of folds

Possible improvement: averaging over several CV random partitions

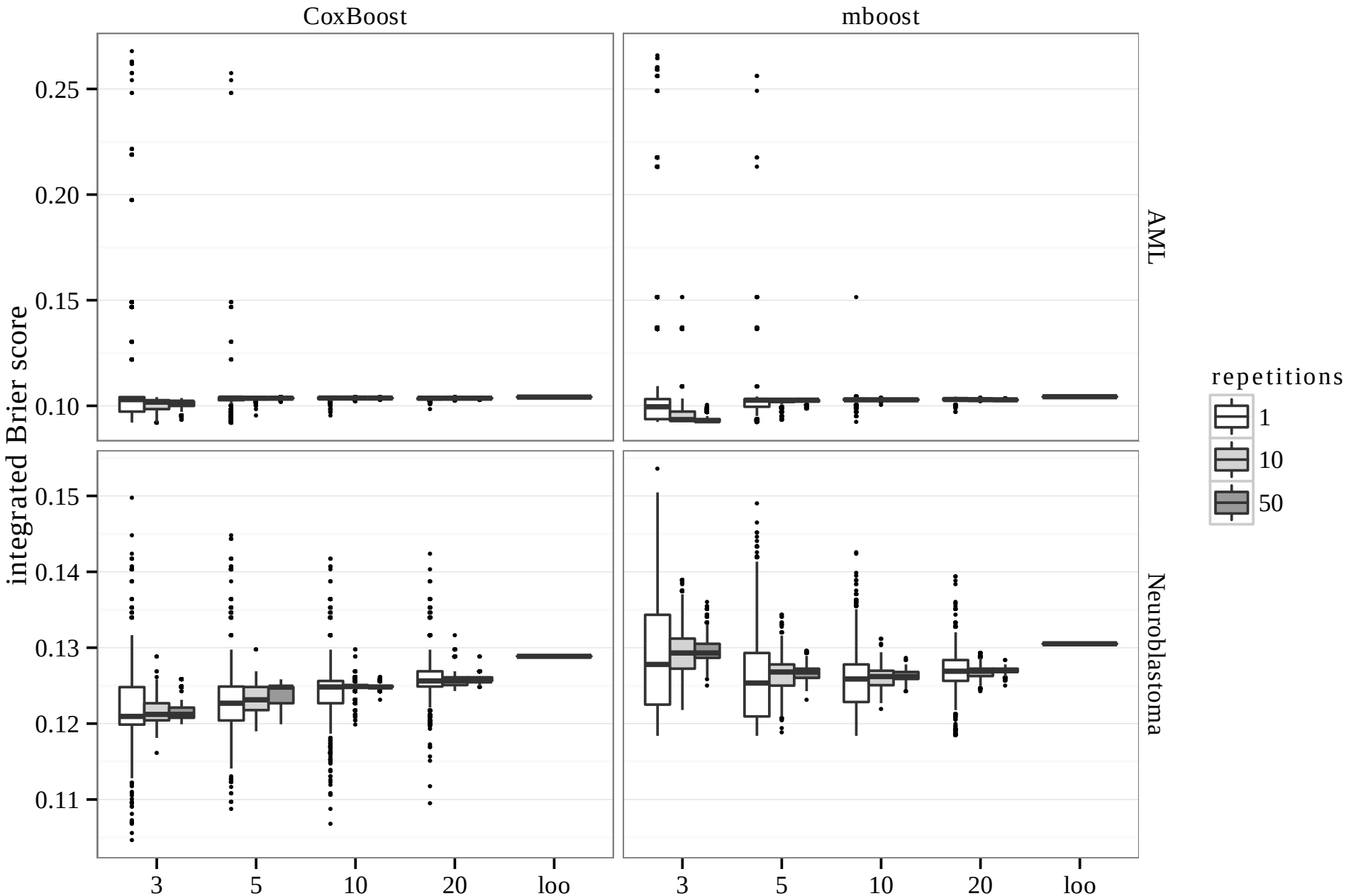
- Idea: averaging over Q K-fold-CV random partitions

$$\overline{cvpl}(m) = \frac{1}{Q} \sum_{q=1}^{Q} cvpl_q(m)$$

- choose $m_{stop} = \arg\max_m \overline{cvpl}(m)$

$\rightarrow$ Reducing the variance on the choice of m_{stop}

$\rightarrow$ More computational time


CoxBoost
mboost
AML
Neuroblastoma
integrated Brier score
0.25
0.20
0.15
0.10
0.15
0.14
0.13
0.12
0.11
3
5
10
20
loo
repetitions
1
10
50

Conclusion / Outlook

- Higher number of folds leads to lower variability in prediction accuracy
- Repeated CV leads to lower variability in prediction accuracy
- The more repetitions, the lower the variability

  $\Rightarrow$ converges to leave-p-out estimator (full CV), see Fuchs et al. (2013)

References I

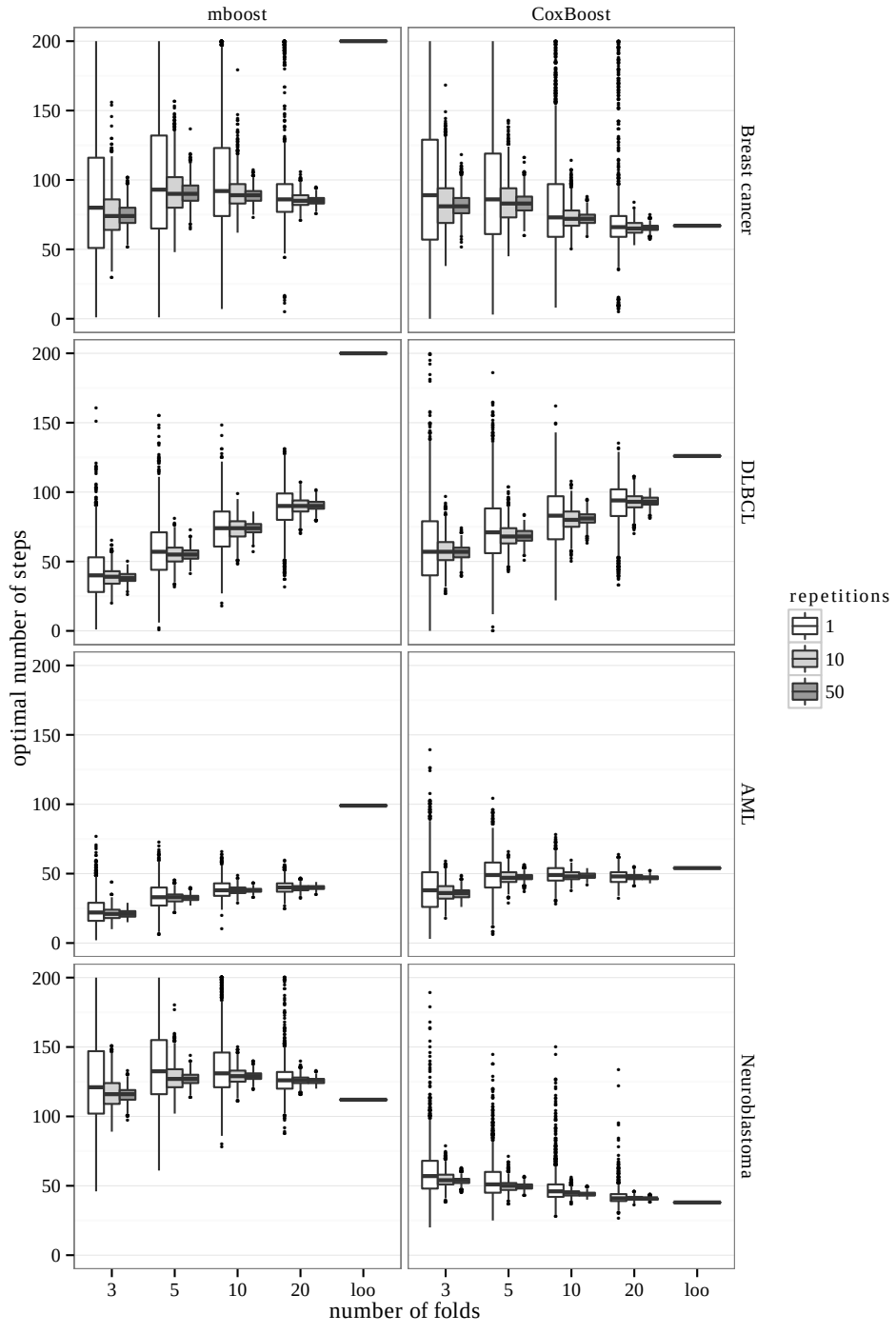
- Binder, Harald and Martin Schumacher (2008). "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models". In: *BMC bioinformatics* 9.1, p. 14.
- Fuchs, Mathias et al. (2013). *A U-statistic estimator for the variance of resampling-based error estimators*. URL: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-17654-2.
- Hothorn, Torsten et al. (2006). "Survival ensembles". In: *Biostatistics* 7.3, pp. 355–373.
- Metzeler, Klaus H et al. (2008). "An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia". In: *Blood* 112.10, pp. 4193–4201.
- Oberthuer, André et al. (2008). "Subclassification and individual survival time prediction from gene expression data of neuroblastoma patients by using CASPAR". In: *Clinical Cancer Research* 14.20, pp. 6590–6601.

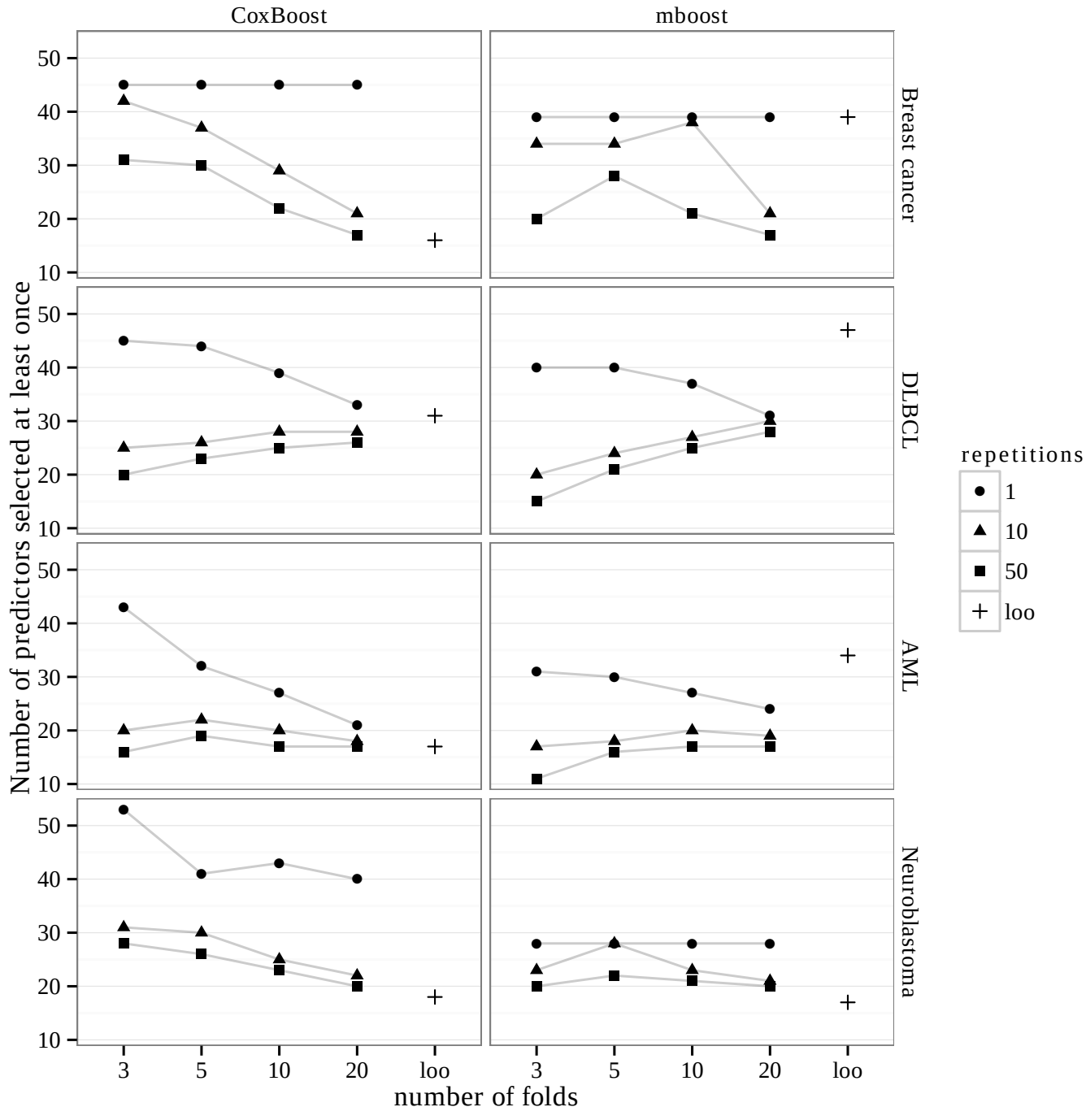
Thank you!

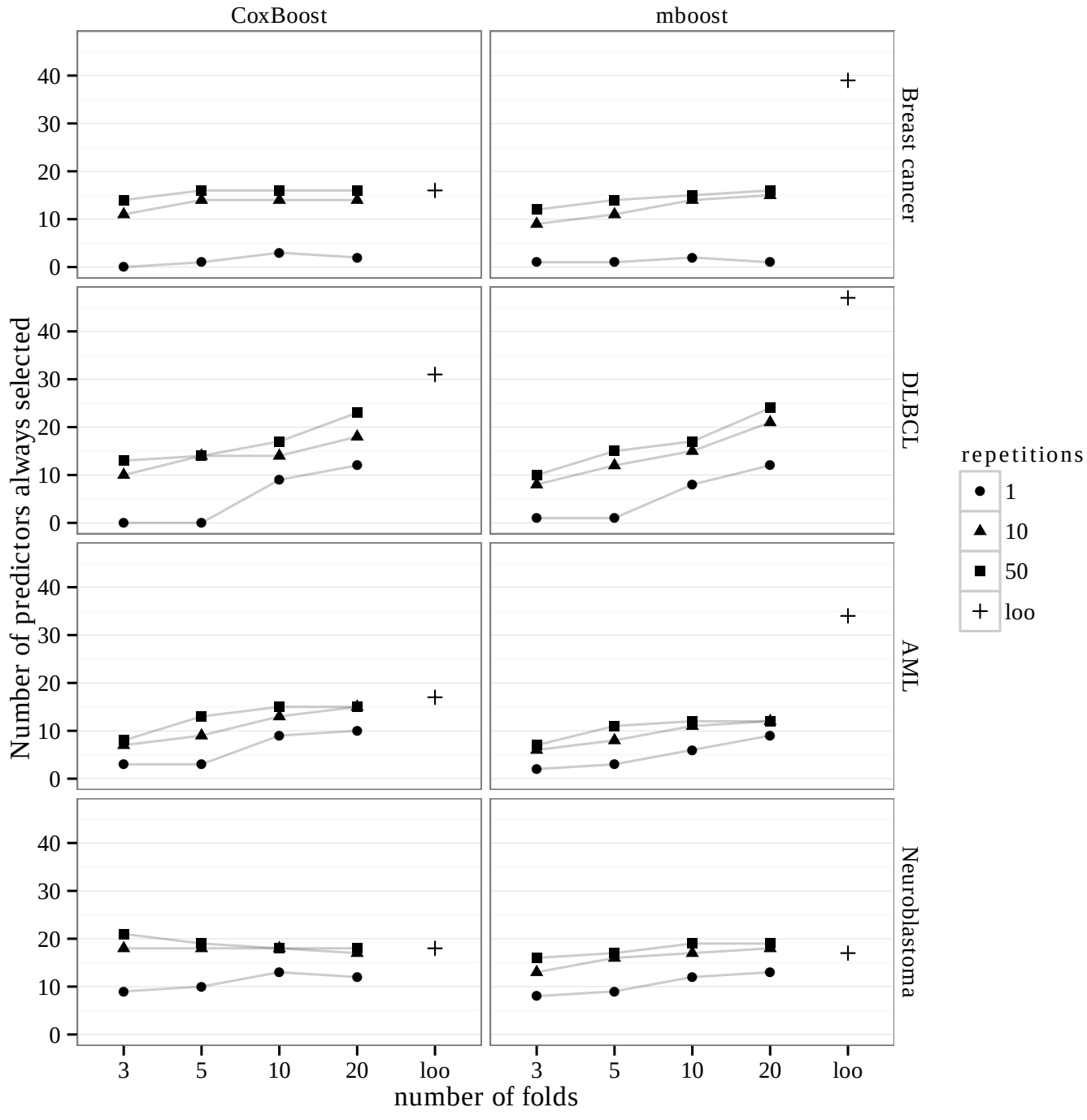
heidi.seibold@uzh.ch

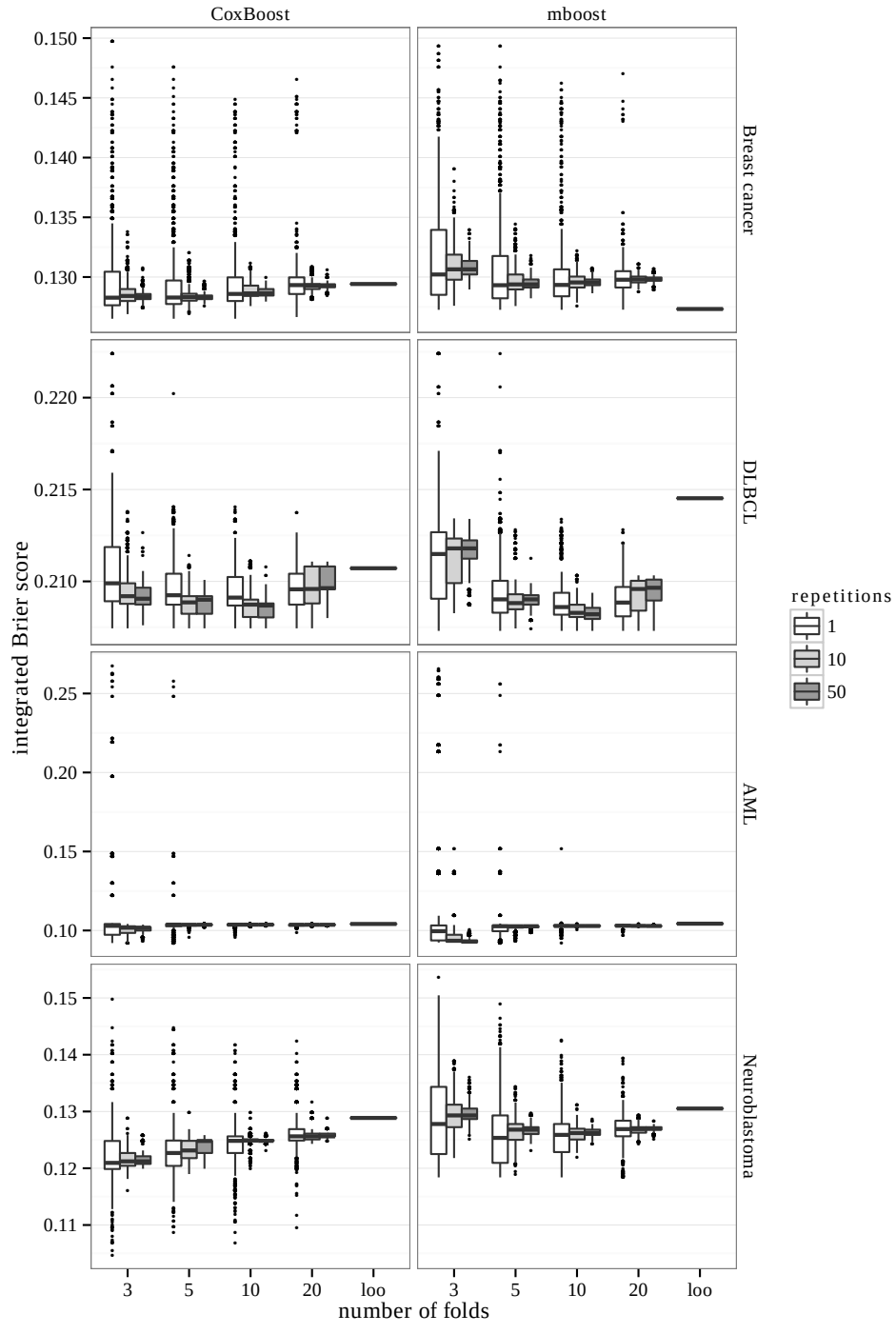
Data sets: training and validation set

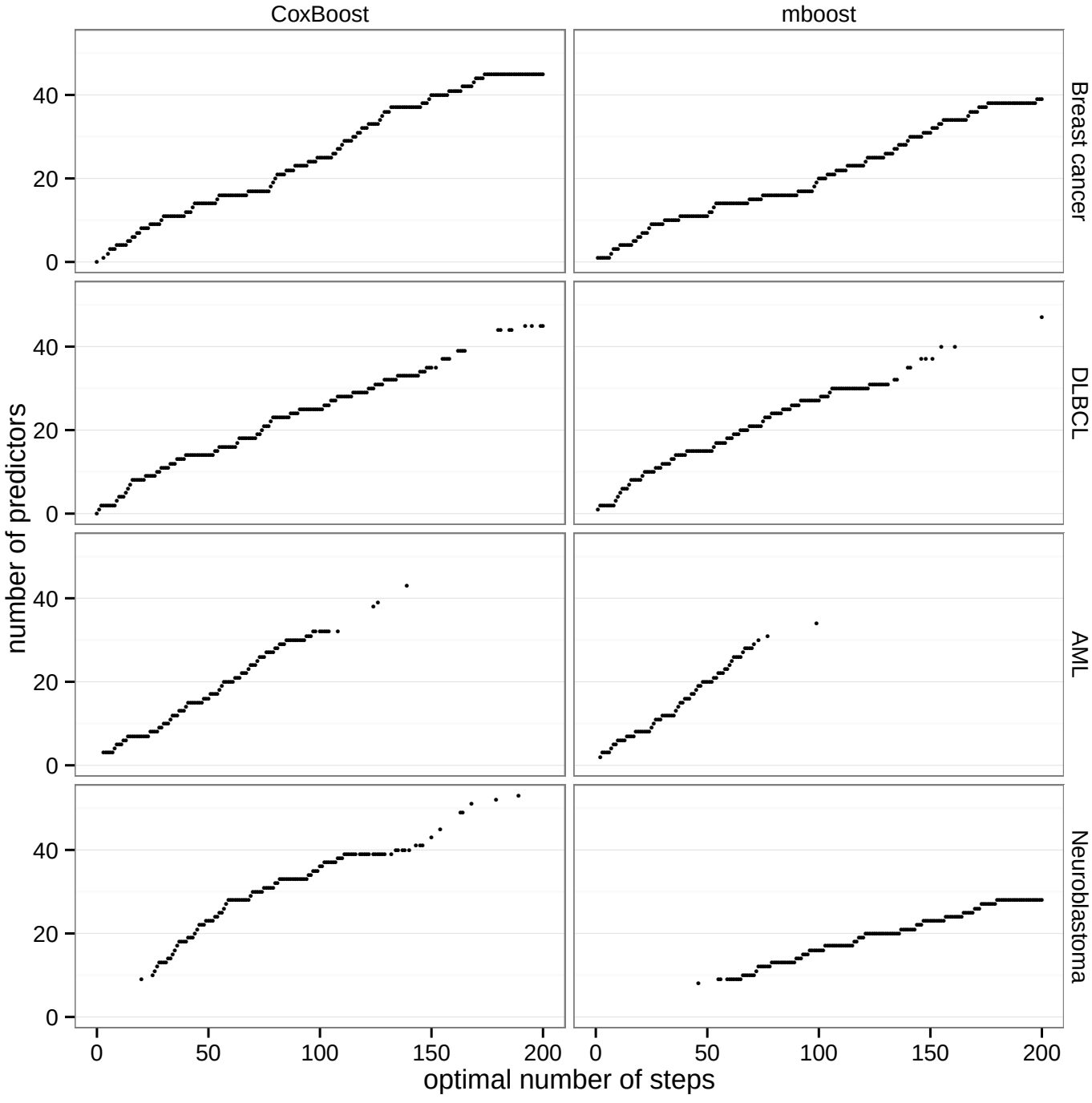
- Metzeler et al. (2008) :
  training: patients enrolled in trial between 1999 and 2003
  validation: patients enrolled in trial in 2004
- Oberthuer et al. (2008):
  training: patients of the German Neuroblastoma Trials
  valitation: patients from centers in other countries

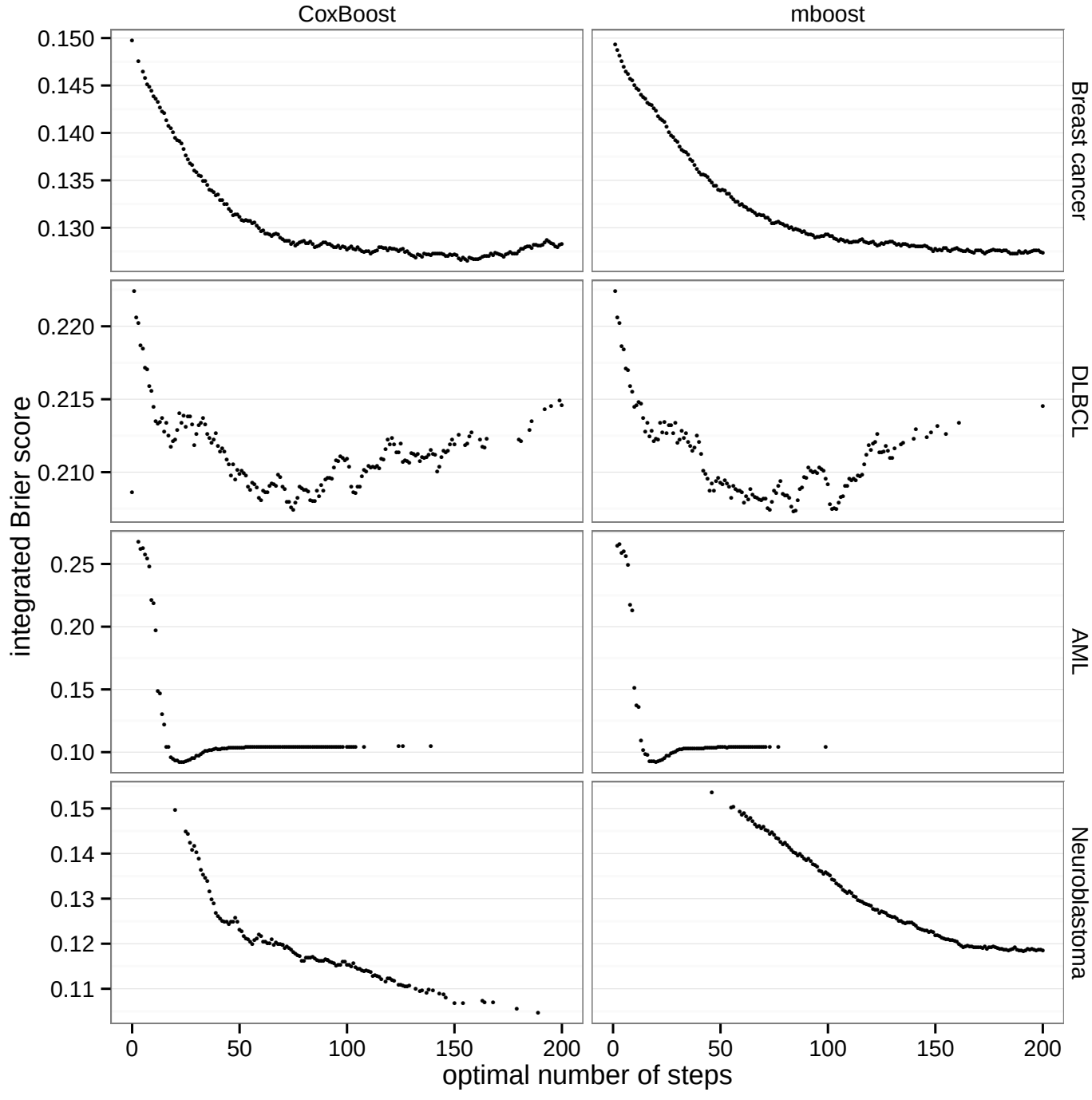

mboost
CoxBoost
Breast cancer
DLBCL
AML
Neuroblastoma
optimal number of steps
number of folds
3
5
10
20
loo
0
50
100
150
200
repetitions
1
10
50


CoxBoost
mboost
Breast cancer
DLBCL
AML
Neuroblastoma
Number of predictors selected at least once
number of folds
3
5
10
20
loo
10
20
30
40
50
repetitions
1
10
50
loo


CoxBoost
mboost
Breast cancer
DLBCL
AML
Neuroblastoma
Number of predictors always selected
number of folds
3
5
10
20
loo
0
10
20
30
40
repetitions
1
10
50
loo


CoxBoost
mboost
Breast cancer
DLBCL
AML
Neuroblastoma
integrated Brier score
number of folds
3
5
10
20
loo
repetitions
1
10
50


CoxBoost
mboost
Breast cancer
DLBCL
AML
Neuroblastoma
number of predictors
optimal number of steps
0
20
40
0
50
100
150
200


CoxBoost
mboost
Breast cancer
DLBCL
AML
Neuroblastoma
integrated Brier score
optimal number of steps
0.150
0.145
0.140
0.135
0.130
0.220
0.215
0.210
0.25
0.20
0.15
0.10
0.15
0.14
0.13
0.12
0.11
0
50
100
150
200