Appendix

In this section we describe the procedure used to build our machine learning model.

Derivation of the machine learning model

We used an ensemble of classifiers to achieve a low variance model. From the derivation cohort, data is randomly split to extract 80% for training (derivation train set) and 20% for testing (derivation test set). This is done by subsampling without replacement. This procedure is repeated 400 times to generate 400 random subsamples (or training/test pairs). The training sets were used to estimate an ensemble of classifiers while the test sets were used to assess the performance of these classifiers (mean Area under ROC curve and 95% CI).

For each training set subsample, a classification model was estimated using the derivation train set. Estimation of the classifier contains two phases: feature selection and classifier design. In *feature selection*, we used an established statistical technique - a generalized linear model with l_1-norm and l_2-norm penalty (alpha parameter set to 0.1 and lambda parameter selected using 5-fold internal cross-validation) [1]. Features with nonzero coefficients were selected. Next, using this feature set, the parameters of a *linear Support Vector machine* [2] classifier were estimated. For SVM implementation, we used the open source package LIBSVM [3].

The above procedure generates an ensemble of 400 classifiers to be tested against on the held-out validation cohort. Three such classifier-ensembles were built, one for each survival prediction tasks (i.e. prediction at 6, 12 and 24 months periods).

Predictors for the machine learning models

Table 1 EMR-based predictors

demographics

- gender
- age
- spoken language
- country of origin
- religion
- occupation
- marital status
- insurance type

cancer specific diagnoses

- primary site
- tumor stream (e.g., breast)
- tumor
- morphology code
- topology code

patient history (in the previous 1 month, 3 months, and 6 months)

- number of inpatient admissions
- number of ED visits
- number of admissions from ED
- longest length of hospital stay
- average length of hospital stay
- number of operations
- number of oncology visits
- number of histology tests
- discharge diagnoses in ICD-10
- diagnosis-related groups codes
- procedure codes

Table 2 ECO-based predictors.

patient demographics

Gender
Age

tumour characteristics

primary site (in ICD-10 code)
tumour stream
morphology (in ICD-O-3 code)
histologic grade
metastatic sites
most valid basis of diagnosis
performance status diagnosis
stage basis (pathological or clinical)
stage (TNM)
tumour size
nodes taken
positive nodes

breast cancer related variables

oestrogen receptor
progesterone receptor
human epidermal growth factor receptor 2 (HER2)

References

1. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996;**58**(1):267-88
2. Cortes C, Vapnik V. Support vector machine. Machine learning 1995;**20**(3):273-97
3. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2011;**2**(3):27