Table E2: Technical details of data processing method

E2.1 Data pre-processing

- Duplicated entries and non-patient (test) rows were removed from the merged data fields. Identifiers were cross-checked for consistency between the data sources.
- There was one instance of gender incorrectly entered. This was rectified using other protected health information (PHI) fields. All PHI fields were thereafter deleted from the merged data.
- Heights, weights and body mass index (BMI) was checked for each patient. Typographic errors in height were corrected using height and BMI. Similarly, incorrect weights were rectified using height and BMI.

E2.2 Missing value imputation

- Nicotine use was missing in 20% of patients. This parameter was excluded from further analysis without imputation.
- Missing values in domicile status and reasons for pre-radiotherapy pain could not be safely imputed from the remaining parameters. These two parameters were excluded from analysis.
- Eight patients had none of weight, height and BMI recorded. These were imputed using multivariate imputation (MICE) with predictive means matching. Each BMI was then recomputed. Heights and weights were excluded from analysis.
- Chemotherapy status was missing in 2 cases. Imputation based on six nearest-neighbours in age, gender, tumour location and lung cancer type was used to estimate the likely values. By inspection, these cases were all NSCLC patients of about median age, therefore it was deemed reasonable to accept the imputed assignment of "concurrent chemotherapy".
- Patients had been scanned using a mixture of 3- and 4-D CT. Absolute volumes of primary tumours and involved nodes in 4D-CT were imputed from 3D-CT, and vice versa. The 3- and 4-D CT volumes were strongly linearly correlated with each other.
- The status of pre-radiotherapy pain medication in 3 cases was manually imputed manually by re-checking medication notes. None of these 3 patients actually had any record of pre-existing pain medication, therefore the pre-radiotherapy pain status was assigned as negative.
- Nurse-scored acute esophagitis was not retrievable in 3 cases.

E2.2 Parameter selection

- Parameters related to PMO strictly post facto were excluded from the model. Each of these related to medication prescribed as a result of PMO, therefore these were not potentially predictive parameters.
- Parameters with zero variance were excluded from the model.
- Highly-correlated clinical parameters (Pearson coefficient > 0.85) to existing parameters were excluded from the model.
- Calendar dates, treatment plan identifiers and free text fields were also excluded from the model.