Developing Infrared Spectroscopic Detection for Stratifying Brain Tumour Patients:

Glioblastoma Multiforme *vs.* Lymphoma

James M. Cameron[1], Holly J. Butler[2], Benjamin R. Smith[2], Mark G. Hegarty[2], Michael D. Jenkinson[3],

Khaja Syed[4], Paul M. Brennan[5], Katherine Ashton[6], Timothy Dawson[6], David S. Palmer[2,7], Matthew J.

Baker[1,2]

Supplementary Information

Table S1 – Retrospective 724 patient cohort information

	Brain Cancer	*Non-Cancer*
Total	487	237
Sex (M/F)	280/207	149/84
Age Range	21-96	19-69
Average Age	61	35

Table S2 - Retrospective 724 patient cohort breakdown

	Tumour Type	**WHO Grade**	**Total**
Cancer	Glioblastoma multiforme	IV	260
	Gliosarcoma	IV	4
	Oligodendroglioma	II	11
	Diffuse astrocytoma	II	23
	Anaplastic astrocytoma	III	10
	Oligoastrocytoma	II	3
	Glioma	I	7
	Pilocytic astrocytoma	I	9
	Pleomorphic xanthoastrocytoma	II	1
	Schwannoma	I	14
	Ependymoma	II	6
	Hemangiopericytoma	II/III	2
	Haemangioblastoma	I	1
	Ganglioglioma	I	1
	Medulloblastoma	IV	1
	PPTID	II/III	1
	Meningioma	I	46
	Pituitary adenoma		29
	Lymphoma		2
	Metastasis		56
Control			237
		Total	724

Table S3 - GBM v lymphoma patient cohort information

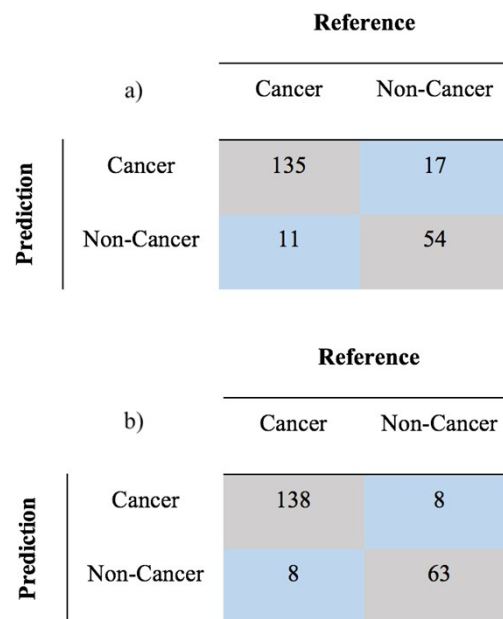|	*GBM*	*Lymphoma*
Total	71	41
Sex (M/F)	48/23	27/14
Age Range	30-85	27-86
Average Age	61	60

Figure S1 - Confusion matrices for the test set of the 724 dataset (a) initial random forest (RF) model, and (b) RF with SMOTE. Grey represents correct predictions, and blue represents incorrect predictions

Reference

	Cancer	Non-Cancer
Cancer	140	13
Non-Cancer	6	58

a) — Prediction

Reference

	Cancer	Non-Cancer
Cancer	130	6
Non-Cancer	16	65

b) — Prediction

Figure S2 - Confusion matrices for the test set of the 724 dataset (a) initial partial least squares-discriminant analysis (PLS-DA) model, and (b) PLS-DA with SMOTE. Grey represents correct predictions, and blue represents incorrect predictions

Reference

	Cancer	Non-Cancer
Cancer	136	13
Non-Cancer	10	58

a) — Prediction

Reference

	Cancer	Non-Cancer
Cancer	134	7
Non-Cancer	12	64

b) — Prediction

Figure S3 - Confusion matrices for the test set of the 724 dataset (a) initial support vector machine (SVM) model, and (b) SVM with SMOTE. Grey represents correct predictions, and blue represents incorrect predictions
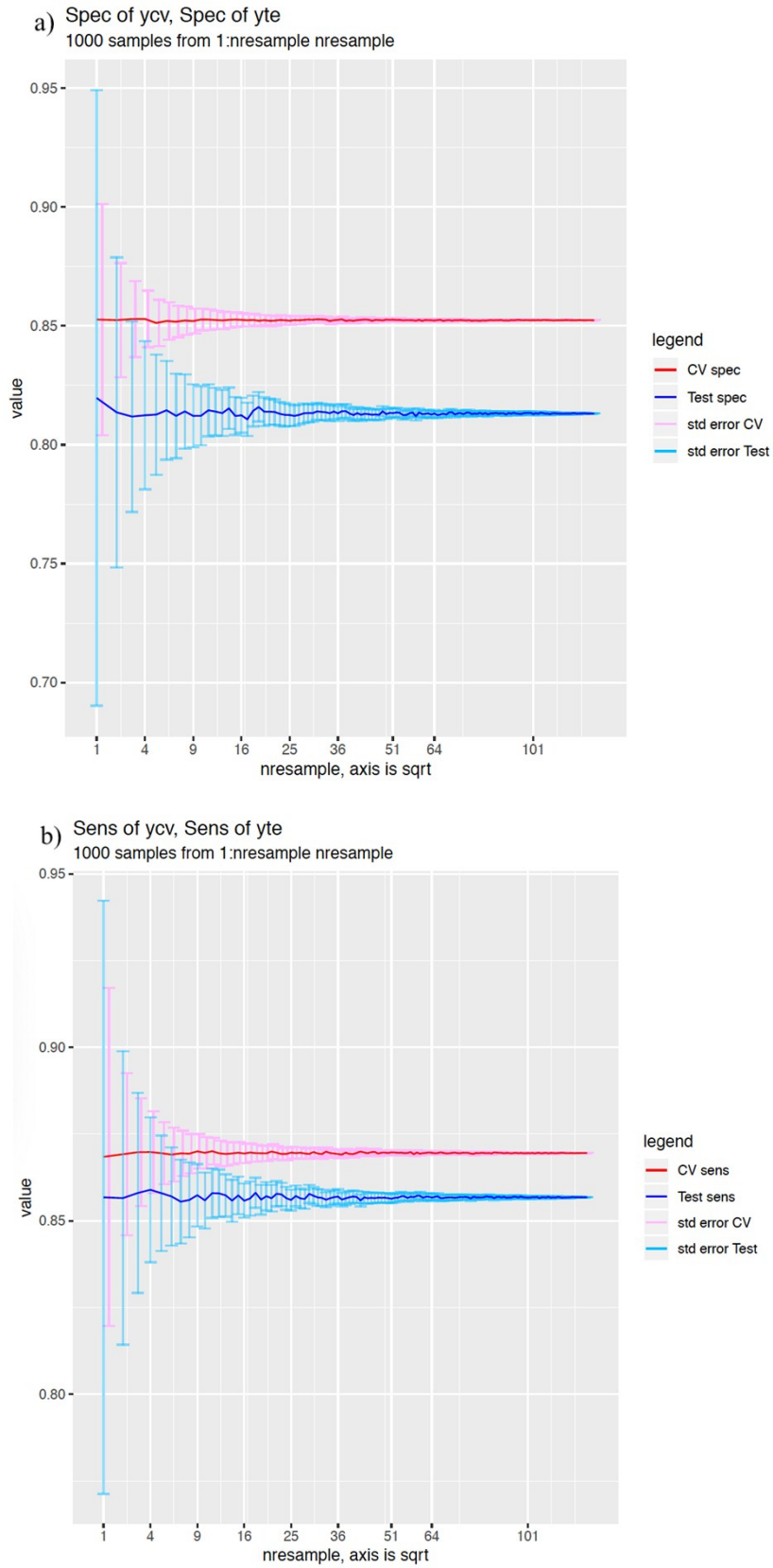
Figure S4 - Bootstrapping analysis to determine sufficient number of resamples required for the lymphoma vs glioblastoma patient dataset: (a) the sensitivity and (b) specificity

		Reference	
		GBM	Lymphoma
Prediction	GBM	22	2
	Lymphoma	2	11

Figure S5 - Confusion matrices for the test set of the glioblastoma v lymphoma dataset; partial least squares-discriminant analysis model with SMOTE. Grey represents correct predictions, and blue represents incorrect predictions