ARTICLE OPEN

Epidemiology

Cancer Grade Model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer

E. Amiri Souri[1], A. Chenoweth[2,3,4], A. Cheung[2,3,4], S. N. Karagiannis[2,3,4] and S. Tsoka[1]✉

BACKGROUND: Prognostic stratification of breast cancers remains a challenge to improve clinical decision making. We employ machine learning on breast cancer transcriptomics from multiple studies to link the expression of specific genes to histological grade and classify tumours into a more or less aggressive prognostic type.
MATERIALS AND METHODS: Microarray data of 5031 untreated breast tumours spanning 33 published datasets and corresponding clinical data were integrated. A machine learning model based on gradient boosted trees was trained on histological grade-1 and grade-3 samples. The resulting predictive model (Cancer Grade Model, CGM) was applied on samples of grade-2 and unknown-grade (3029) for prognostic risk classification.
RESULTS: A 70-gene signature for assessing clinical risk was identified and was shown to be 90% accurate when tested on known histological-grade samples. The predictive framework was validated through survival analysis and showed robust prognostic performance. CGM was cross-referenced with existing genomic tests and demonstrated the competitive predictive power of tumour risk.
CONCLUSIONS: CGM is able to classify tumours into better-defined prognostic categories without employing information on tumour size, stage, or subgroups. The model offers means to improve prognosis and support the clinical decision and precision treatments, thereby potentially contributing to preventing underdiagnosis of high-risk tumours and minimising over-treatment of low-risk disease.

British Journal of Cancer (2021) 125:748–758; https://doi.org/10.1038/s41416-021-01455-1

BACKGROUND

Despite progress in early detection and personalised targeted therapy, breast cancer remains a major cause of fatality and quality of life reduction worldwide [1]. As breast cancer encompasses a heterogeneous group of diseases, precision diagnosis and treatment mandates accurate tumour stratification into clinically distinct subgroups [2]. Classifying tumours based on intrinsic features like histological grade or subtype can predict disease behaviour more accurately than time-dependent prognostic factors such as tumour size and stage [1, 3, 4], and thereby provide better insight into prognosis and suitable treatment strategies [5, 6].

A significant advance has been achieved in determining treatment on the basis of subtypes of breast cancer, which are immunohistochemically classified based on the expression of oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) [7, 8]. ER-positive cancers can benefit from hormone therapies such as tamoxifen or aromatase inhibitors, while the monoclonal antibody trastuzumab is typically used as an adjuvant therapy for HER2 breast cancer in combination with chemotherapy. Patients with triple-negative breast cancer (TNBC), whose tumours lack ER, PR and HER2 expression, do not benefit from the current development of targeted therapies, with treatment options mostly relying on primary surgery, radiotherapy and chemotherapy cocktail [9]. Four further molecular subtypes (luminal A, luminal B, HER2-enriched and basal-like) are identified by PAM50 classification [10] based on the expression of 50-gene signatures, with most basal-like carcinomas, which are usually triple negative, having aggressive phenotype and high relapse rates. Despite significant progress in understanding disease subtypes, the challenge of matching patient clinical characteristics and tissue molecular patterns to prognosis or to a therapeutic strategy remains pertinent [11, 12]. Predictive models of molecular profiling are urgently needed to prevent the underdiagnosis of high-risk tumours and to minimise the over-treatment of low-risk disease, which may help reduce the need for aggressive systemic therapies [13, 14].

Histological grade is a well-described prognostic factor, reflecting tumour morphological characteristics and clinical behaviour of the disease [4, 5]. For instance, in a process of systematic treatment selection, prompt consideration of neoadjuvant or adjuvant chemotherapy is needed for grade-3 tumours, while

[1]Department of Informatics, Faculty of Natural and Mathematical Sciences, King's College London, London, UK. [2]St. John's Institute of Dermatology, School of Basic and Medical Biosciences, King's College London, London, UK. [3]NIHR Biomedical Research Centre at Guy's and St. Thomas' Hospitals and King's College London, London, UK. [4]Breast Cancer Now Research Unit, School of Cancer and Pharmaceutical Sciences, Guy's Cancer Centres, King's College London, London, UK. ✉email: sophia.tsoka@kcl.ac.uk

Received: 15 November 2020 Revised: 29 April 2021 Accepted: 28 May 2021
Published online: 15 June 2021

patients with grade 1 could benefit from long-term follow-up [15]. However, for 30–60% of breast tumours diagnosed as grade 2, treatment may be difficult to assign, as they represent an intermediary and highly variable state in morphology, underlying biology and risk of distant metastasis recurrence [4]. Therefore, patients diagnosed with these tumours are at risk of under-treatment or over-treatment [14]. It has been suggested that only grades 1 and 3 be used towards treatment choice [16] and that grade 2 are not informative [3] without additional metrics [15]. Therefore, accurate stratification of grade-2 tissues poses significant challenges.

The key to resolving these challenges is the development of integrative, systems-level analyses that can capture the multiple facets of disease while also guiding the search for specific molecular cascades that discriminate between disease phenotypes [17]. Analysis of high-throughput gene expression [18] profiles in cancer tissues through computational methods with predictive capabilities [19, 20], such as machine learning models, is critical [13, 21]. Here, we report a bioinformatics strategy where transcriptomic profiles across multiple datasets were integrated and a machine learning model was generated to classify tumours into relevant histological grades. The resulting Cancer Grade Model (CGM) was then used to dissect the molecular subtypes present in grade-2 and unknown-grade cancers and re-classify them into grade-1-like (low-risk) or grade-3-like (high-risk) categories. By interpreting the classification model, key genes were extracted to predict metastasis, risk of relapse and overall survival (OS), regardless of traditional histologically defined receptor status. These markers might also provide potential therapeutic targets for the disease currently lacking treatment options. We report the application of gradient boosted trees on a large dataset of samples integrated from multiple breast cancer studies; however, it is important to note that this strategy can be applied to other types of high-throughput data or cancer types in the future.

MATERIALS AND METHODS

Computational framework

Figure 1 shows the overall computational framework, which includes data pre-processing and integration, development of the CGM machine learning model and cross-validation, as well as key steps of model interpretation through feature prioritisation, prognostic data analysis and pathway enrichment.

Dataset, pre-processing and integration

Gene expression data from 33 breast cancer datasets corresponding to platforms GPL570 [Genome U133 Plus 2.0] and GPL96 [Genome U133A] were obtained from Gene Expression Omnibus (Table 1) [22]. Samples with prior treatment were excluded (data selection workflow in Supplementary Figure S1). A total of 5031 tumour samples and 70 normal samples were collected along with their clinical characteristics, including ER, PR, and HER2 status, distant metastasis-free survival (DMFS), relapse-free survival (RFS), OS and PAM50 subtype [10] (Supplementary Table S1).

Data integration comprised normalising raw intensity data in Affymetrix CEL data files through Robust Multichip Average [23] and removing batch effects with COMBAT [24] using R 3.3 and related libraries. Probes were mapped to genes, and in cases of multiple probes mapping to one gene, the average value was taken. The data processing pipeline is illustrated in Fig. 1a. For the implementation of the machine learning model and cross-validation, samples corresponding to grade 1 and grade 3 (henceforth termed *development-dataset*, 2002 samples) were used to build the classification model, which was later applied on grade-2 and unknown-grade samples (termed *prediction-dataset*, 3029 samples) to stratify them into low- and high-risk categories.

Machine learning model development

The model was formulated as binary classification on the development-dataset, where gene expression values represented input variables and cancer grade 1 vs. grade 3 were output variables. An overview of the machine learning pipeline is shown in Fig. 1b. Model performance metrics were calculated through 10-fold cross-validation repeated ten times to prevent bias on the splitting dataset. The development-dataset was split into 80% training set, 10% validation set for hyperparameter tuning and developing the classifier, and 10% test set for evaluating performance metrics. Detecting outliers in train and validation datasets was performed with the *K*-nearest neighbour algorithm [25] of PyOD [26] library was used by measuring the distance of an observation to *k*th nearest neighbour as the outlying score. The relevant samples were removed from the training set. In cases of imbalanced data, over-sampling through Synthetic Minority Over-sampling Technique (SMOTE) [27] was used on the imbalanced dataset [13]. The machine learning model was built on XGBoost [28] (eXtreme Gradient Boosting), a machine learning method that combines weak learners (decision trees) to achieve stronger class discrimination.

Grid search was performed on training set samples within each cross-validation fold to find the best set of hyperparameters. The best-performing model was selected (termed CGM, with hyperparameters of maximum tree depth = 5, subsample ratio = 0.6, minimum child weight = 1, and gamma = 0.5 in XGBoost [28]) and applied to classify the prediction-dataset samples into high or low risk. For selecting genes that were most important in classification, the Gain metric [29] was used to calculate the average across all cross-validation sets. The smallest set of genes was identified, which reflected the highest Gain value and kept the model performance at the same level as and when all genes were employed. In addition, the SHAP [30] (SHapley Additive exPlanations) method was used to obtain features globally important for classification. Python 3.7.3, Scikit-learn 0.21.2 and XGBoost 0.90 were used to implement the models. Principal component analysis (PCA) [31] was applied to visualise the grouping of samples based on the expression of selected genes.

Analysis and interpretation of machine learning prediction results

Clinical parameters (time and event of DMFS, RFS and OS) were used to evaluate differences between high- and low-risk groups in the development and prediction-dataset. Survival analysis was undertaken using Kaplan–Meier (KM) methodology [32]. For multivariable analysis, Cox's proportional hazard model by CoxPHFitter [33] was used. Stratification was cross-referenced with PAM50 [10] and genomic tests for risk of metastasis and relapse (OncotypeDx [34], EndoPredict [35] and Gene expression Grade Index (GGI) developed in genefu [36] package in Bioconductor). Pathway enrichment was performed by mapping the selected biomarkers on Reactome [37] with *p* values calculated on a hypergeometric model [38] and a cut-off of 0.05. An overview is illustrated in Fig. 1c.

RESULTS

Model training and risk prediction of grade-2 and unknown-grade samples

Processed gene expression datasets used in this study were derived through the integration of 5031 samples (429 grade 1, 1409 grade 2, 1573 grade 3 and 1620 unknown grade) and 12,806 genes (Fig. 1a) spanning 33 breast cancer studies (Table 1). The dataset was divided into a *development-dataset* comprising the grade-1 and grade-3 samples (total of 2002) with which the machine learning model CGM was trained and tested, and a *prediction-dataset* of grade-2 and unknown-grade samples (total of 3029) for prediction and classification of these samples into low risk (1130) or high risk (1899). Clinical parameters and survival properties of the development and prediction datasets were used to validate the model. A machine learning model based on gradient boosted trees was built on grade-1 and grade-3 samples of the development-dataset (Fig. 1b), with a performance metric of 89% accuracy (Fig. 2a). Use of the Gain metric yielded a prioritisation list of gene features in terms of importance in the classification of grade-1 vs. grade-3 tumours. The 70 top genes with the highest Gain value (Supplementary Table S2) that kept model accuracy at the same level as all genes (Supplementary Figure S2) were selected and used for classification with model performance maintaining high accuracy of 90% (Fig. 2a).

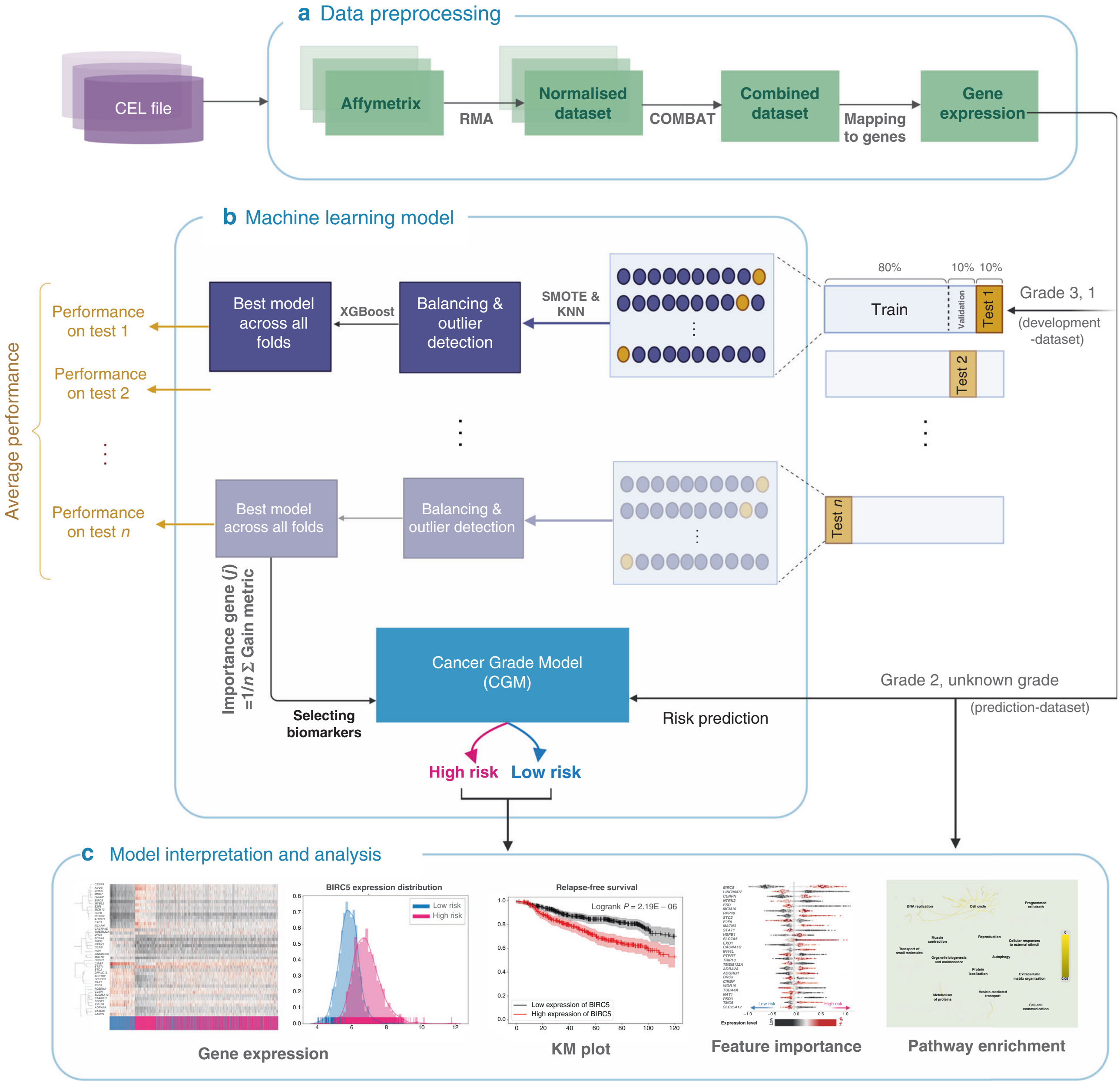


Fig. 1 CGM overall computational framework. a Data pre-processing: input data comprise CEL files from different studies. For each dataset the Affymetrix probe intensity was normalised, datasets merged and batch effects removed. **b** Schematic representation of the CGM machine learning model. The XGBoost library was trained on grade-1 and grade-3 cancer samples over ten-fold cross-validation, which was repeated ten times. Feature importance based on average gain value was employed to reduce gene features to 70 markers with the best discriminatory properties. CGM was then applied to stratify grade-2 and unknown-grade samples into more or less aggressive cancer phenotypes (high-risk or low-risk cohorts, respectively). **c** Interpretation of machine learning model results. A series of analyses employing prognostic features, such as patient survival, gene importance based on SHAP values and pathway enrichment, were used to evaluate biomarkers and the stratification of cancer samples achieved by CGM.

For samples across all histological grades (including samples of unknown grade), PCA based on the expression of the 70 selected genes (Fig. 2b) illustrated that grade-3 samples separated well from milder phenotypes in grade 1, whereas grade-2 tumours reflected a widely diverging histological manifestation [3, 4]. Analysis of prognostic factors confirmed that grade-1 samples had substantially better survival outcome compared to grade 3, with grade 2 in mid-way (Supplementary Figure S3a). Clinical subtype information of the samples is shown in Fig. 2c. Most samples in high-grade disease (grade 3) were ER−, PR− and HER2+, while ER+, PR+ and HER2− groups were linked to low-grade samples (grade 1) mostly (Fig. 2c). There were also more grade-3 samples in the HER2-enriched and basal-like groups according to PAM50 molecular subtypes, compared with the less aggressive luminal A and luminal B samples. Importantly, in terms of both immunohistological status and PAM50 subtype, a large proportion of samples were grade 2 or unknown grade, which highlights the clinical challenge of assigning the right treatment for this group of patients and identifying means to dissect this diverse cancer cohort into more informative risk groups.

After building the XGBoost classifier on grade-1 and grade-3 samples, the CGM predictive framework was used to assign risk to

Table 1. List of GEO datasets employed in this study.

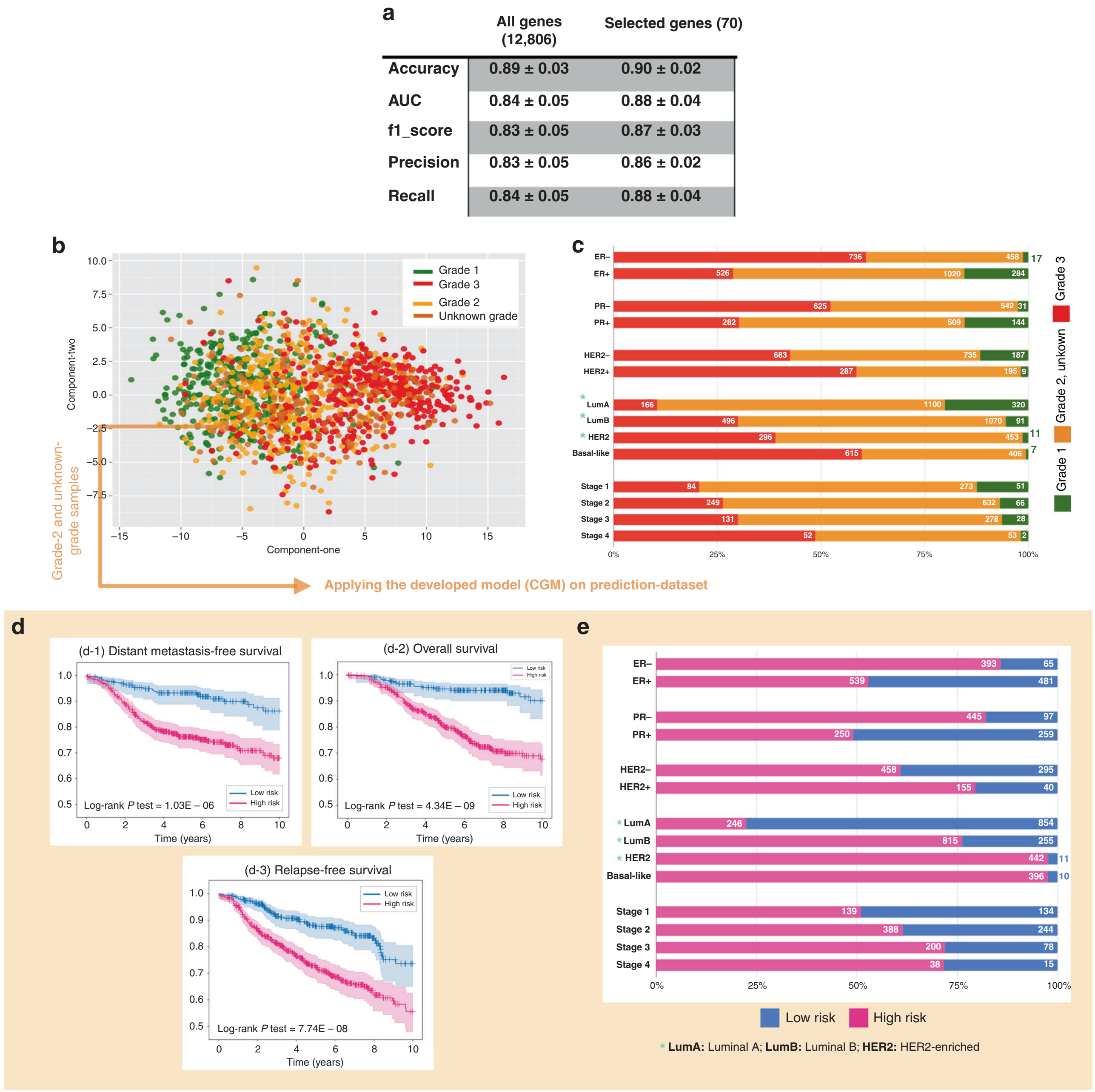
GEO ID	Platform	Sample type	# sample	# grade 1	# grade 2	# grade 3	# NA
GSE11121	GPL96	Primary tumour	200	29	136	35	0
GSE18864	GPL570	Tumour biopsy	84	10	16	58	0
GSE20711	GPL570	Primary tumour	88	13	5	70	0
GSE23593	GPL570	Primary tumour	50	2	23	25	0
GSE27120	GPL570	Primary tumour	28	3	11	14	0
GSE32646	GPL570	Tumour biopsy	115	16	78	21	0
GSE36771	GPL570	Unknown	107	11	42	54	0
GSE42568	GPL570	Primary tumour	104	11	40	53	0
GSE50948	GPL570	Primary tumour	154	0	68	86	0
GSE5460	GPL570	Primary tumour	129	27	32	70	0
GSE11001	GPL570	Primary tumour	30	4	13	13	0
GSE87007	GPL570	Primary tumour	31	5	3	23	0
GSE88770	GPL570	Primary tumour	117	13	96	7	1
GSE7390	GPL96	Primary tumour	198	30	83	83	2
GSE78958	GPL570	Primary tumour	424	88	156	178	2
GSE45255	GPL96	Primary tumour	139	17	52	67	3
GSE61304	GPL570	Tumour adjacent epithelium	62	5	16	37	4
GSE63471	GPL570	Tumour biopsy	142	5	52	81	4
GSE21653	GPL570	Primary tumour	266	45	89	125	7
GSE26639	GPL570	Tumour biopsy	226	15	83	121	7
GSE17907	GPL570	Primary tumour	55	3	10	34	8
GSE10810	GPL570	Primary tumour	32	2	10	10	10
GSE25066	GPL96	Primary tumour biopsy	508	32	180	259	37
GSE47109	GPL570	Primary tumour	246	43	115	49	39
GSE95700	GPL570	Primary tumour	57	0	0	0	57
GSE5327	GPL96	Primary tumour	58	0	0	0	58
GSE48390	GPL570	Primary tumour	81	0	0	0	81
GSE58984	GPL570	Unknown	94	0	0	0	94
GSE103091	GPL570	Primary tumour	104	0	0	0	104
GSE45827	GPL570	Primary tumour	130	0	0	0	130
GSE65194	GPL570	Primary tumour	130	0	0	0	130
GSE1456	GPL96	Primary Ttumour	159	0	0	0	159
GSE102484	GPL570	Primary tumour	683	0	0	0	683

samples in the prediction-dataset (grade-2/unknown-grade group, 3029 samples), thereby re-classifying these samples into high-risk (grade-3-like samples, 1899) or low-risk (grade-1-like samples, 1130). KM analysis for these groups (Fig. 2d) illustrated that RFS, DMFS and OS are significantly better in the low-risk group than in the high-risk group (log-rank *P* 1.03E − 06, 4.34E − 09 and 7.74E − 08, respectively). For instance, within 5 years, 5% of low-risk patients were reported to develop metastasis compared to >20% in the high-risk group (Fig. 2d–1). In terms of immunohistochemical subtypes (Fig. 2e), of samples classified as low risk in the prediction-dataset, 47% were ER+, 51% samples were PR+ and 40% HER2−, thereby representing cases where cytotoxic regimens can be avoided. According to the PAM50 subtype, of samples classified as high risk, 22% were luminal A, 76% luminal B, 98% HER2-enriched and 98% basal-like, linking them to cases benefiting from further systemic treatment.

Multivariable analysis was performed based on several clinical variables (hormone receptors, stage, age, etc.). The analysis showed the association between CGM, tumour stage and ER status with RFS (Supplementary Figure S4a). Therefore, we examined the joint distribution of stage, ER status and CGM. ER− status was generally associated with a high-risk group (grade-3 and CGM high-risk group, see Fig. 2c, e); however, ER-positive status was associated with a heterogeneous mixture of high- and low-risk groups, which CGM can stratify into better prognostic groups with log-rank *P* test = 1.95E − 10 (see Supplementary Figure S4c-1). Similarly, for stage 2 and stage 3, which represent intermediate risk, CGM can define prognostically relevant subgroups (log-rank *P* test = 1.36E − 03, and 1.36E − 03) (see Supplementary Figure S4c-2, 3).

Evaluation of prognostically relevant biomarker genes

The machine learning model involved feature prioritisation through the use of the Gain value (Supplementary Table S2) and the 70 top-scoring genes were further evaluated through hierarchical clustering (Fig. 3a). Expression profile clusters showed association to sample risk labels, which are indicated by the grouping of samples (colour-coded columns according to predicted risk). A similar heatmap of gene expression that also included normal samples is shown in Supplementary Figure S5, which shows that low-risk tumours have not undergone significant gene expression changes when compared to normal breast tissues. Further prioritisation of genes was employed based on SHAP values [30] to determine the contribution of each gene feature in the predictions generated by the CGM framework



	All genes (12,806)	Selected genes (70)
Accuracy	0.89 ± 0.03	0.90 ± 0.02
AUC	0.84 ± 0.05	0.88 ± 0.04
f1_score	0.83 ± 0.05	0.87 ± 0.03
Precision	0.83 ± 0.05	0.86 ± 0.02
Recall	0.84 ± 0.05	0.88 ± 0.04

Fig. 2 Model training and risk prediction. **a** Performance of the CGM methodology in classifying grade-1 and grade-3 samples. **b** Principal component analysis of samples across all three histological grades based on 70 biomarkers. **c** Hormone receptor characteristics and PAM50 subtype status of all samples. **d** Analysis through Kaplan–Meier plots across various survival metrics: distant metastasis-free survival (d-1), recurrence-free survival (d-2) and overall survival (d-3) for patients in the prediction-dataset (grade 2 and unknown grade) classified as either high risk (pink) or low risk (blue) by CGM. **e** Hormone receptor characteristics and PAM50 subtype status in high- and low-risk group samples.

(Fig. 3b). It is noted that (i) genes higher in the list reflect larger overall contribution to the prediction model, (ii) the colour of each data point indicates the expression value of the gene feature in the corresponding sample and (iii) the horizontal position of data points reflect the impact in prediction (i.e. high negative SHAP values show a stronger contribution to the prediction of low risk, whereas high positive values reflect the prediction of high risk). The same analysis employing hierarchical clustering and SHAP values for the development-dataset (grade-1 and grade-3 tumours) is shown in Supplementary Figure S3b, c.

Interpreting the prioritisation list in Fig. 3b can indicate cases of genes overexpressed in the high-risk group or overexpressed in low-risk samples that may be associated with molecules with important cancer-related activity. Mean expression values in the high- and low-risk groups for all 70 selected markers are shown in Supplementary Table S3. As an example, *BIRC5* (overexpressed in high-risk samples) and *LINC00472* (overexpressed in low-risk) are also shown to have significant prognostic value based on RFS calculated across our dataset and KM plotter (Fig. 3c, d). This observation is consistent with known roles for these genes, where high expression of *LINC00472* [39] is known to activate p53 signalling pathway that inhibits cancer development [40], while *BIRC5* encoding Survivin is involved in carcinogenesis by influencing cell division and proliferation and inhibiting apoptosis

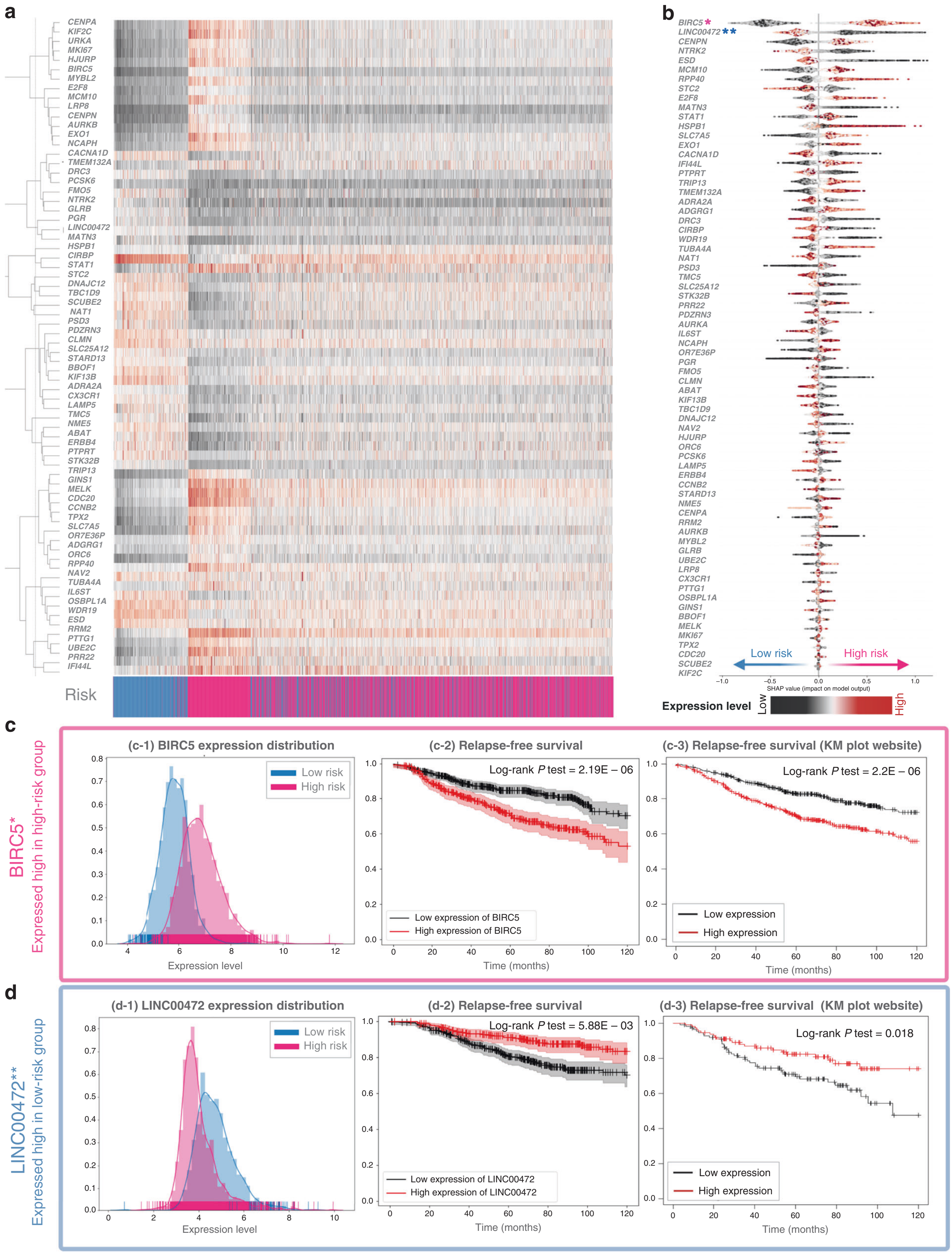


Fig. 3 Analysis of the development dataset. a Hierarchical clustering of biomarker expression across the prediction-dataset (grade 2 and unknown grade). Samples are labelled according to risk assigned by CGM (pink for high risk and blue for low risk). **b** Prioritisation list according to SHAP values for the selected biomarker genes. Colour (black to red) indicates the values of expression of each gene feature, and the horizontal position of each point reflects the impact of that value in the prediction of risk (see text). **c**, **d** *BIRC5* and *LINC00472* as high- and low-risk genes, respectively (c-1, d-1). Expression value distribution for BIRC5 and *LINC00472* and in high- (pink) and low- (blue) risk groups. (c-2, d-2) *BIRC5* and *LINC00472* KM plots for relapse-free survival. (c-3, d-3) *BIRC5* and *LINC00472* KM plots for relapse-free survival are created by the KM plotter website based on grade-2 samples >120 months.

[41]. For other genes, *AURKA*, *PTTG1*, *CDC20*, *SLC7A5*, *E2F8*, *TPX2*, and *TUBA4A*, high expression in the high-risk group was linked to tumour growth and metastasis. On the other hand, increasing levels of *NME5* and *CACNA1D* expression could suppress malignant behaviour. Some of these biomarkers, including *E2F8* [42], *TPX2* [43] and *CACNA1D* [44], have been independently confirmed as tumourigenic or tumour-suppressive in breast cancer, and can thus point towards novel targets for treatment. Employing the Connectivity Map (CMap) [45] to search for potential drugs tested against the 70 marker genes indicated cases where some of these genes were explored as targets in clinical trials (details in Supplementary Table S4), demonstrating the potential of our findings.

Further validation of the 70 biomarkers indicated a significant association with survival outcome (survival calculation details in Supplementary Tables S5 and S6). Pathway enrichment analysis using the Reactome database indicated cell cycle regulation, gene expression and DNA replication as the most affected pathways for the high-risk group (Supplementary Figure S6a,c). Perturbation in any of these pathways was not observed in the low-risk group, indicating slow tumour growth and low metastatic potentials (Supplementary Figure S6b).

Genomic tests for clinical assessment of breast cancer

Results of CGM on re-classifying grade-2 and unknown-grade tumours were compared to currently available genomic tests, namely OncotypeDX [34], EndoPredict [35], and GGI [4]. These tests associate gene activity level to cancer aggressiveness and are recommended in national and international guidelines for proposing adjuvant systemic therapy [6]. Venn diagram (Fig. 4a) showing overlapping biomarkers identified by CGM and the other reported genes by the three methods [4, 34, 35] (full gene list in Supplementary Table S7).

Pairwise comparison of CGM with each of the three methods (Fig. 4b and Supplementary Table S8) showed 91% similarity of sample classification with OncotypeDX (without considering samples predicted in medium-risk group by OncotypeDX), followed by EndoPredict and GGI with 76% and 74% similarity, respectively. It is noted that as GGI does not work on unknown-grade samples, those samples were eliminated for comparison with GGI. We assessed survival through KM plots for DMFS and RFS on the prediction-dataset with all four methods and showed that CGM was more discriminative in prognosis and better in classifying samples into the high- or low-risk groups, as indicated by the relevant P values (Fig. 4c and Supplementary Figure S7). We also performed the multivariable analysis of the prognostic methods (GGI, OncotypeDX and EndoPredict) based on time to relapse using fitting Cox's proportional hazard model in Supplementary Figure S4b, which shows the strongest association of CGM with RFS (hazard ratio (HR) = 1.71, 95% confidence interval (CI) = 1.2–2.43; $P < 0.005$).

DISCUSSION

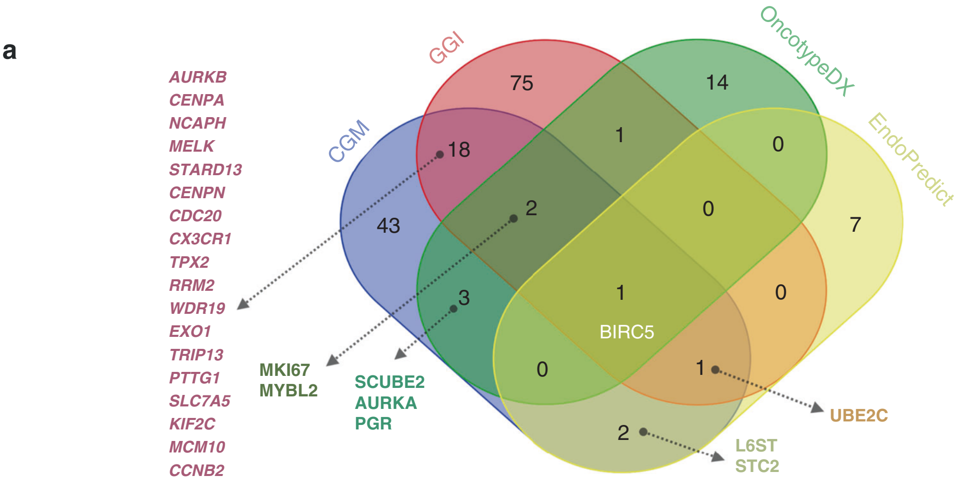
Breast cancer is a heterogeneous disease with significant variance in genetic predisposition and phenotypic characteristics. Despite progress, assigning a more accurate prognosis requires optimisation and significant challenges remain in attaching appropriate treatment to relevant patient subgroups. It is widely accepted that tumour classification based on intrinsic features such as histological grade can predict prognostic features or treatment options more accurately than time-dependent factors [1, 3]. A case in point is the prescription of chemotherapy in grade-3 tumours, but not in those of grade 1. However, for grade-2 phenotypes that represent a heterogeneous cancer class, prognosis and treatment remain particularly challenging with either over-treatment or missed diagnoses being common. To help address these limitations, here we report CGM a machine learning platform based on gradient boosted trees to classify grade-2 and unknown-grade tumours into high and low risk, after training the model on genomic data of high- and low-grade cancers and generating a 70-gene signature. Our methodology includes the use of feature importance metrics such as Gain and SHAP values, which offer means of attributing biological significance to specific genes and are particularly important in counter-balancing the black-box nature of machine learning models.

Based on the 70 genes selected as markers, the CGM prediction model can successfully separate grade-2 and unknown-grade samples into high- or low-risk groups, regardless of the conventional surface receptor immunohistochemistry-based subtypes or the PAM50 molecular-based subgrouping. Basal-like and HER2-enriched subtypes are high-risk tumours, and, in the absence of treatment, these patients have a poor prognosis [46]. In the clinic, the treatment strategy for HER2-enriched breast cancers is mainly by targeted therapy with trastuzumab in combination with chemotherapy, while basal-like/TNBC are treated with radiotherapy and chemotherapy until the recent development of olaparib for BRCA-mutated TNBC, and anti-PD-L1 immunotherapy in combination with chemotherapy for advanced disease [47, 48]. In our predictive model, 98% of HER2-enriched and 98% of basal-like grade 2/unknown-grade tumours are classified as high-risk and would require chemotherapy, in line with current disease control settings for these two tumour types. However, prediction via our CGM model would benefit patients in groups classified as less aggressive.

We evaluated the prognostic power of the CGM model and found that among patients with grade-2/unknown-grade cancer, about half of those with ER/PR tumour expression and 22% of patients with luminal A breast cancer have high-risk disease. The CGM model can determine the potential prognosis of patients based on their genetic signature and can indicate whether a patient with grade 2 or unknown grade has a high-risk disease and thus may require immediate treatment intervention or whether they have a low-risk disease and may benefit from less aggressive treatment strategies [49]. In clinical practice, our model is unlikely to be applied to very aggressive subtypes such as HER2-overexpressed and TNBC, but to subtypes that have a more "intermediate" prognosis such as luminal B. Treatment of luminal B cancers is typically based on an in-depth list of criteria, including size, lymph node involvement, grade, Ki67 status and a low personal risk of relapse determined by other gene expression signature models such as OncotypeDX [50]. It is possible that our CGM model may be used to determine the potential prognosis of these patients, and thus inform treatment strategy independent of clinical characteristics.

Our model also revealed genes that may contribute towards cancer progression. Some of these are known to be associated with breast cancer, while others are less well understood and would benefit from further functional characterisation. Literature search and pathway analysis using Reactome [51] identified key functions and pathways associated with our biomarkers. These mostly associated with the cell cycle, DNA replication, transcription, and signal transduction. Furthermore, while the high-risk gene set involved multiple genes connecting the above pathways, the low-risk gene set had no pathways with more than one gene affected (Supplementary Figure S6). This indicates that those samples feature low cancer proliferation rates, lack of escape mechanisms and lower metastatic potential, and could hence be classified as low-risk cancers that might not require systemic and cytotoxic therapies.

Although unsurprisingly the most fundamental traits identified for a high-risk group were predominantly related to the dysregulation of cell cycle checkpoints and transcription regulator molecules [52, 53], individual genes could also provide insights for novel treatment development. Downregulation of *BIRC5* (survivin), which is the top-ranked biomarker revealed by CGM, was reported



b

Grade 2/ unknown grade	#Low risk	#High risk	Description
GGI	835	574	1620 samples do not have grade information (GGI cannot be applied to them)
OncotypeDX	668	1564	797 samples are predicted ad medium-risk group
Endopredict	395	2634	–
CGM	1130	1899	–

c

	Relapse-free survival (RFS) log-rank *P* test	Distant metastasis-free survival (DMFS) log-rank *P* test
GGI	1.23E – 05	1.13E – 03
OncotypeDX	9.94E – 06 (high vs. low), 1.22E – 01 (low vs. medium), 1.15E – 03 (medium vs. high)	1.27E – 05 (high vs. low), 7.94E – 02 (low vs. medium), 3.34E – 04 (medium vs. high)
Endopredict	6.24E – 03	1.29E – 04
CGM	7.74E – 08	1.03E – 06

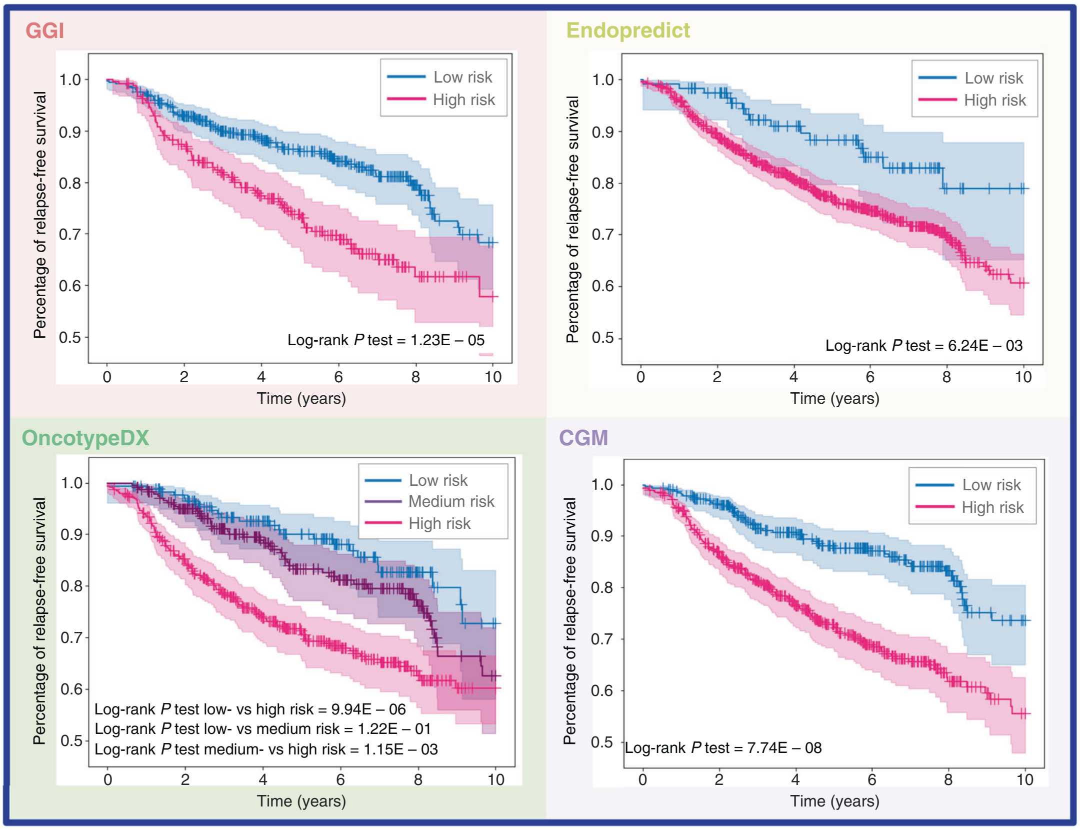


Fig. 4 Comparison of CGM with existing genomic tests. **a** Venn diagram for comparison of gene biomarkers identified via CGM with other genomic test methods, namely GGI, OncotypeDX and Endopredict. **b** The number of samples in prediction-dataset (grade 2 and unknown grade) stratified to relevant risk groups through GGI, OncotypeDX, Endopredict and CGM. **c** Comparing GGI, OncotypeDX, Endopredict and CGM risk prediction on log-rank *P* test on distant metastasis- and relapse-free survival. Survival curves for relevant cohort predictions by each method are also shown.

as an inhibitor of tumour cell migration and invasion through the PI3K/Akt signalling pathway [54]. Survivin remains a promising target for drug discovery and breast cancer therapeutics, ranging from selective suppressants that disrupt survivin function [55] to antisense oligonucleotides that degrade survivin messenger RNA [56]. Survivin peptide-mediated immunotherapy has also been tested in clinical trials [57]. An interaction between *AURKA* (Aurora Kinase A) and MAPK pathway has been proposed for a new treatment strategy using a combination of AURKA and MEK1/2 inhibitors in breast cancer [58, 59]. PTTG1 contributes to different cancer-promoting pathways that can increase cell growth through a nuclear exclusion of p27 [60]. CDC20 is overexpressed in TNBC and could be used as a treatment target [61], while the expression of SLC7A5 part of the large neutral amino acid transporter small subunit 1 heterodimer has been reported to correlate with luminal cancers [62], and anti-SLC7A5 targeted therapies have been developed for endocrine therapy-resistant tumours [63]. On the other hand, low expression of LINC00472, thought to act as a tumour suppressor, has been reported to suppress nuclear factor-κB signalling [39], which contributes to tumour progression and metastasis [64]. NME5 plays a key role in DNA proofreading and repair [65] and would be predicted to be associated with low-risk cancers.

We further compared CGM with well-known diagnostic methods in the literature [6] to validate the predictive power of our model. OncotypeDX and EndoPredict predict distant recurrence of early-stage breast cancer (stage 1 and 2) based on a set of 21 and 11 genes, respectively (Supplementary Table S7). OncotypeDX [34] calculates a recurrence score between 0 and 100 to reflect the likelihood of breast cancer recurrence within 10 years and to classify patients into groups of low (<26), medium (26–30) and high (>30) score, where higher scores indicate a greater likelihood of recurrence. EndoPredict [35] analyses tumour gene activation to provide a risk score for ER/PR+, HER2− samples as either low risk or high risk, and recurring as distant metastasis within 10 years. Finally, GGI [4] employs 97 differentially expressed genes between histologic grade-1 and grade-3 tumours, which are selected through the analysis of 189 breast cancer microarray datasets to classify grade-2 tumours into two groups (grade 1 or grade 3) to suggest relevant treatment.

Unlike CGM, OncotypeDX classifies patients into three groups (low risk, medium risk or high risk), which leads to the assignment of samples into a non-informative medium-risk group category, while the difference between low- and medium-risk groups is not significant (Fig. 4c). Unlike CGM, GGI requires the grade of the tumours to be known a priori, which may not always be available and relies on parameters calculated for the given dataset (its scale and offset), thus making it dependent on the dataset used. Furthermore, GGI ignores the importance of genes in cancer by assigning equal weight to all of them and relying on their sum. This can lead to a lack of interpretability and may prevent understanding of the individual contribution of each gene in aggressive tumours. EndoPredict does not employ a medium group; however, it tends to place more patients into the high-risk group (about 7 times more than low risk, Fig. 4b). There are about twice more patients in the low-risk groups in the two other methods (668 by OncotypeDX and 835 by GGI) compared to EndoPredict (395), suggesting that EndoPredict tends to a larger number of high-risk predictions (395 low risk and 2634 high risk), thereby leading to over-treatment.

In overview, moderately differentiated tumours represented by those diagnosed as grade 2 are particularly difficult to treat, leading to over- or under-treatment in this patient group. Genomic information can identify novel predictive biomarkers and signalling pathways indicative of disease progression or phenotype. In this study, we present a computational pipeline using gradient boosted trees to analyse large and complex datasets, integrated from multiple breast cancer studies, to discover patient subtypes and derive an understanding of prognosis. Even though we exemplify our platform on gene expression data for histological grade prediction, our strategy is generic and can be applied to other types of high-throughput data and clinical labels.

Our model can assign high- and low-risk groups, without using clinical data such as tumour size, stage or breast cancer subgroup information, offering a potential means to avoiding underdiagnosis of high-risk tumours and minimise over-treatment of low-risk diseases, thus helping to reduce unnecessary toxic treatments. In addition, our findings of key genes differentiating patient phenotypes may point to important regulators of aggressive disease phenotype and lead to a better understanding of underlying malignant disease mechanisms across subgroups. These could ultimately point to novel therapeutic targets applicable to specific disease types. Even though CGM is implemented and evaluated on breast cancer, capitalising on the rich and well-studied datasets for this cancer, we stress that our methodology can be used in other types of cancer where subtypes are less well understood, and the application of well-established machine learning methods would add valuable prognostic insights.

DATA AVAILABILITY

The datasets that support the findings of this study are collected from the Gene Expression Omnibus (GEO) repository [http://www.ncbi.nlm.nih.gov/geo].

CODE AVAILABILITY

The CGM code will be made available at: https://github.com/elmira-amiri/CGM.

REFERENCES

1. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT. et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30:1194–220.
2. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nat Rev Dis Prim. 2019;5:66.
3. Wang M, Klevebring D, Lindberg J, Czene K, Grönberg H, Rantalainen M. Determining breast cancer histological grade from RNA-sequencing data. Breast Cancer Res. 2016;18:48.
4. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006;98:262–72.
5. Daveau C, Baulies S, Lalloum M, Bollet M, Sigal-Zafrani B, Sastre X, et al. Histological grade concordance between diagnostic core biopsy and corresponding surgical specimen in HR-positive/HER2-negative breast carcinoma. Br J Cancer. 2014;110:2195–2200.
6. Harris LN, Ismaila N, McShane LM, Andre F, Collyar DE, Gonzalez-Angulo AM, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline. J Clin Oncol. 2016;34:1134–50.
7. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.
8. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA. 2001;98:10869–74.
9. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121:2750–67.
10. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
11. Jahn SW, Bösl A, Tsybrovskyy O, Gruber-Rossipal C, Helfgott R, Fitzal F, et al. Clinically high-risk breast cancer displays markedly discordant molecular risk predictions between the MammaPrint and EndoPredict tests. Br J Cancer. 2020;122:1744–6.
12. Yersal O, Barutca S. Biological subtypes of breast cancer: prognostic and therapeutic implications. World J Clin Oncol. 2014;5:412–24.
13. Ming C, Viassolo V, Probst-Hensch N, Dinov ID, Chappuis PO, Katapodi MC Machine learning-based lifetime breast cancer risk reclassification compared with

the BOADICEA model: impact on screening recommendations. Br J Cancer. 2020. https://doi.org/10.1038/s41416-020-0937-0.

14. Matikas A, Foukakis T, Swain S, Bergh J. Avoiding over- and undertreatment in patients with resected node-positive breast cancer with the use of gene expression signatures: are we there yet? Ann Oncol. 2019;30:1044–50.
15. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010;12:207.
16. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thürlimann B, Senn H-J. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2009. Ann Oncol. 2009;20:1319–29.
17. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. BMC Cancer. 2019;19:230.
18. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
19. Bertucci F, Finetti P, Ostrowski J, Kim WK, Kim H, Pantaleo MA, et al. Genomic Grade Index predicts postoperative clinical outcome of GIST. Br J Cancer. 2012;107:1433–41.
20. Christos S, Lajos P. Gene-expression signatures in breast cancer. N Engl J Med. 2009;11: 790–800. https://doi.org/10.1056/NEJMra0801289.
21. Halama N. Machine learning for tissue diagnostics in oncology: brave new world. Br J Cancer. 2019;121:431–3.
22. Fajarda O, Duarte-Pereira S, Silva RM, Oliveira JL. Merging microarray studies to identify a common gene expression signature to several structural heart diseases. BioData Min. 2020;13:8.
23. Irizarry RA. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249–64.
24. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
25. Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. SIGMOD Rec. 2000; 29:427–438. https://doi.org/10.1145/335191.335437.
26. Zha Y, Nasrullah Z, Li Z. PyOD: A Python Toolbox for Scalable Outlier Detection. J Mach Learn Res. 2019;20:1–7. https://arxiv.org/abs/1901.01588.
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57.
28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–94. https://arxiv.org/abs/1603.02754.
29. Tibshirani R, Hastie T, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Vol. 1. Springer series in statistics. New York: Springer; 2001.
30. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. Long Beach: Curran Associates; Long Beach, CA, USA; 2017. p. 4765–74. https://arxiv.org/abs/1705.07874.
31. Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Stat Soc Ser B. 1999;61:611–22.
32. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457–81.
33. Cameron D-P, Jonas K, Noah J, Sean R, Ben K, Paul Z, et al. CamDavidsonPilon/lifelines: 0.25.10. Zenodo. 2021. https://doi.org/10.5281/zenodo.4579431.
34. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.
35. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res. 2011;17:6012–20.
36. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32:1097–9.
37. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12:477–9.
38. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20:3710–5.
39. Wang Z, Katsaros D, Biglia N, Shen Y, Loo L, Yu X, et al. ERα upregulates the expression of long non-coding RNA LINC00472 which suppresses the phosphorylation of NF-κB in breast cancer. Breast Cancer Res Treat. 2019;175: 353–68.
40. Zou A, Liu X, Mai Z, Zhang J, Liu Z, Huang Q, et al. LINC00472 acts as a tumor suppressor in NSCLC through KLLN-mediated p53-microRNA-149-3p and microRNA-4270. Mol Ther. 2019;17:563–77.
41. Li F, Aljahdali I, Ling X. Cancer therapeutics using survivin BIRC5 as a target: what can we do after over two decades of study? J Exp Clin Cancer Res. 2019;38:368.
42. Ye L, Guo L, He Z, Wang X, Lin C, Zhang X, et al. Upregulation of E2F8 promotes cell proliferation and tumorigenicity in breast cancer by modulating G1/S phase transition. Oncotarget. 2016;7:23757–71.
43. Chen M, Zhang H, Zhang G, Zhong A, Ma Q, Kai J, et al. Targeting TPX2 suppresses proliferation and promotes apoptosis via repression of the PI3k/AKT/P21 signaling pathway and activation of p53 pathway in breast cancer. Biochem Biophys Res. Commun. 2018;507:74–82.
44. Phan NN, Wang C-Y, Chen C-F, Sun Z, Lai M-D, Lin Y-C. Voltage-gated calcium channels: novel targets for cancer therapy. Oncol Lett. 2017;14:2059–74.
45. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.
46. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96.
47. Robson M, Im S-A, Senkus E, Xu B, Domchek SM, Masuda N, et al. Olaparib for metastatic breast cancer in patients with a germline *BRCA* mutation. N Engl J Med. 2017;377:523–33.
48. Schmid P, Adams S, Rugo HS, Schneeweiss A, Barrios CH, Iwata H, et al. Atezolizumab and Nab-Paclitaxel in advanced triple-negative breast cancer. N Engl J Med. 2018;379:2108–21.
49. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh I-T, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. Lancet Oncol. 2010;11:55–65. https://doi.org/10.1016/S1470-2045(09)70314-6.
50. Foulon A, Theret P, Rodat-Despoix L, Kischel P. Beyond chemotherapies: recent strategies in breast cancer treatment. Cancers. 2020;12:2634.
51. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2019. https://doi.org/10.1093/nar/gkz1031.
52. F. P, S. V, F. A, C. S. A clinical evaluation of treatments that target cell cycle machinery in breast cancer. Expert Opin Pharmacother. 2019;20:2305–15.
53. Thu KL, Soria-Bretones I, Mak TW, Cescon DW. Targeting the cell cycle in breast cancer: towards the next phase. Cell Cycle. 2018;17:1871–85.
54. Shang X, Liu G, Zhang Y, Tang P, Zhang H, Jiang H, et al. Downregulation of BIRC5 inhibits the migration and invasion of esophageal cancer cells by interacting with the PI3K/Akt signaling pathway. Oncol Lett. 2018. https://doi.org/10.3892/ol.2018.8986.
55. Clemens MR, Gladkov OA, Gartner E, Vladimirov V, Crown J, Steinberg J, et al. Phase II, multicenter, open-label, randomized study of YM155 plus docetaxel as first-line treatment in patients with HER2-negative metastatic breast cancer. Breast Cancer Res Treat. 2015;149:171–9.
56. Tanioka M, Nokihara H, Yamamoto N, Yamada Y, Yamada K, Goto Y, et al. Phase I study of LY2181308, an antisense oligonucleotide against survivin, in patients with advanced solid tumors. Cancer Chemother Pharmacol. 2011;68:505–11.
57. Kameshima H, Tsuruma T, Kutomi G, Shima H, Iwayama Y, Kimura Y, et al. Immunotherapeutic benefit of α-interferon (IFNα) in survivin2B-derived peptide vaccination for advanced pancreatic cancer patients. Cancer Sci. 2013;104:124–9.
58. Eterno V, Zambelli A, Villani L, Tuscano A, Manera S, Spitaleri A, et al. AurkA controls self-renewal of breast cancer-initiating cells promoting wnt3a stabilization through suppression of miR-128. Sci Rep. 2016;6:28436.
59. Gandhi S, Gil M, Khoury T, Takabe K, Puzanov I, Gelman I, et al. Abstract P2-06-17: a novel interaction of AURKA with MAPK pathway in breast cancer cells as a potential therapeutic target. Cancer Res. 2019;79:P2-P2-06–17.
60. Yoon C-H, Kim M-J, Lee H, Kim R-K, Lim E-J, Yoo K-C, et al. *PTTG1* oncogene promotes tumor malignancy via epithelial to mesenchymal transition and expansion of cancer stem cell population. J Biol Chem. 2012;287:19516–27.
61. Cheng S, Castillo V, Sliva D. CDC20 associated with cancer metastasis and novel mushroom-derived CDC20 inhibitors with antimetastatic activity. Int J Oncol. 2019. https://doi.org/10.3892/ijo.2019.4791.
62. El Ansari R, Craze ML, Miligy I, Diez-Rodriguez M, Nolan CC, Ellis IO, et al. The amino acid transporter SLC7A5 confers a poor prognosis in the highly proliferative breast cancer subtypes and is a key therapeutic target in luminal B tumours. Breast Cancer Res. 2018;20:21.
63. Sevigny CM, Sengupta S, Luo Z, Liu X, Hu R, Zhang Z, et al. SLCs contribute to endocrine resistance in breast cancer: role of SLC7A5 (LAT1). Cancer Biol. 2019. https://doi.org/10.1101/555342.

64. DiDonato JA, Mercurio F, Karin M. NF-κB and the link between inflammation and cancer. Immunol Rev. 2012;246:379–400.

65. Puts GS, Leonard MK, Pamidimukkala NV, Snyder DE, Kaetzel DM. Nuclear functions of NME proteins. Lab Invest. 2018;98:211–8.

ACKNOWLEDGEMENTS

We acknowledge support by Breast Cancer Now (147; KCL-BCN-Q3); the Cancer Research UK King's Health Partners Centre at King's College London (C604/A25135); the CRUK/NIHR in England/DoH for Scotland, Wales and Northern Ireland Experimental Cancer Medicine Centre (C10355/A15587). The authors are solely responsible for study design, data collection, analysis, decision to publish and preparation of the manuscript. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

AUTHOR CONTRIBUTIONS

EAS and ST conceived the study design. EAS and AC conducted the data collection. EAS and ST conducted computational experiments. EAS, A Chenoweth, A Cheung, SNK and ST analysed and interpreted the data. EAS, A Chenoweth, A Cheung, SK and ST wrote the initial manuscript drafts. EAS, A Chenoweth, A Cheung, SNK and ST performed critical editing. All authors approved the final version of the manuscript.

FUNDING

This research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) based at Guy's and St. Thomas' NHS Foundation Trust and King's College London (IS-BRC-1215-20006).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41416-021-01455-1.

Correspondence and requests for materials should be addressed to S.T.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.