Machine Learning to Guide the use of Adjuvant Therapies for Breast Cancer

Ahmed Alaa (ahmedmalaa@ucla.edu)
University of California, Los Angeles
Deepti Gurdasani
William Harvey Research Institute, Queen Mary University
Adrian Harris
Department of Oncology, Weatherall Institute of Molecular Medicine, University of Oxford
Jem Rashbass
National Cancer Registration and Analysis Service, Public Health England
Mihaela van der Schaar
University of Cambridge

Article

Keywords: machine learning, breast cancer, adjuvant therapy

Posted Date: August 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-53594/v1

Machine Learning to Guide the use of Adjuvant Therapies for Breast Cancer

Ahmed M. Alaa[1], Deepti Gurdasani[2], Adrian L. Harris[3],
Jem Rashbass[4], and Mihaela van der Schaar[1,5,6,†]

[1] *Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA.*

[2] *William Harvey Research Institute, Queen Mary University, London, UK.*

[3] *Department of Oncology, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK.*

[4] *National Cancer Registration and Analysis Service, Public Health England, London, UK.*

[5] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK.*

[6] *The Alan Turing Institute, London, UK.*

[†] **Corresponding author: Mihaela van der Schaar, email: mv472@cam.ac.uk.**

Abstract

Accurate prediction of the individualized survival benefit of adjuvant therapy is key to making informed therapeutic decisions for patients with early invasive breast cancer. Here, we use a state-of-the-art *automated* and *interpretable* machine learning algorithm to develop a breast cancer prognostication and treatment benefit prediction model — Adjutorium — using data from large-scale cohorts of nearly 1 million women captured in the national cancer registries of the United Kingdom and the United States. We trained and internally validated the Adjutorium model on 395,862 patients from the UK National Cancer Registration and Analysis Service (NCRAS); we then externally validated the model among 571,635 patients from the US Surveillance, Epidemiology, and End Results (SEER) Program. Adjutorium exhibited significantly improved accuracy compared to the major prognostic tool in current clinical use (PREDICT v2.1) in both internal and external validation (AUC-ROC for 5-year survival prediction in NCRAS was 0.835, 95% CI: 0.833–0.837 and 0.755, 95% CI: 0.753–0.757 for Adjutorium and PREDICT v2.1. In SEER, the AUC-ROC performance was 0.815, 95% CI: 0.813–0.817 and 0.775, 95% CI: 0.772–0.778 for Adjutorium and PREDICT v2.1, respectively). Importantly, our model substantially improved accuracy in specific subgroups known to be under-served by existing models. Adjutorium is currently implemented as a web-based decision support tool (vanderschaar-lab.com/adjutorium/) to aid decisions on adjuvant therapy in women with early breast cancer, and can be publicly accessed by patients and clinicians worldwide[1].

[1] The website is currently password protected and the online tool Adjutorium can be activated by entering password 12321 each time it is accessed.

Main

Breast cancer is the most common cancer among women globally, with incidence rates varying from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women in Western Europe.[1, 2] While prognosis of early-stage breast cancer has improved substantially since the introduction of adjuvant endocrine and chemotherapies,[3] these treatments need to be used judiciously, with careful balancing of risks and benefits, particularly in patients' subgroups where their utility is as yet unclear.[4, 5] Over the years, various breast cancer prognostication models have been developed to enable tailored post-surgical therapeutic decisions by predicting the survival profiles of individual patients on the basis of their clinicopathological features. Of these, PREDICT v2.1 (https://predict.nhs.uk) has been the most commonly used worldwide;[6, 7, 8] it was recently endorsed by the American Joint Committee on Cancer (AJCC),[9] was accessed through more than 1 million sessions from 100 cities all over the world in the period spanning from 2011 to 2020 (https://breast.predict.nhs.uk/statistics.html), and is the recommended tool for adjuvant therapy planning in the current NICE guidelines.[10]

However, despite its widespread use, PREDICT v2.1 has been shown to under-perform in specific subgroups of patients, including older patients, patients with tumours over 50mm, small ER-positive tumours, or larger ER negative tumours.[11] Over or under-estimation of the survival rates within specific patient subgroups could lead to under or over-treatment, thereby, negatively impacting patient outcomes.[12, 13, 14, 15] We hypothesize that the limitations of existing tools arise from: (1) the lack of flexibility in the underlying Cox regression method predominantly used to develop prognostic models,[16, 7] and (2) the derivation of models using outdated and relatively modest-sized cohorts where certain subgroups of patients may not be sufficiently represented. Machine learning (ML) technologies that can readily infer complex patterns from data, supported with big data resources provide the opportunity to address the aforementioned limitations.[17, 18]

Here, we use a state-of-the-art automated ML algorithm, AutoPrognosis,[19] to develop and validate *Adjutorium*; a breast cancer prognostication model that predicts patient survival and adjuvant treatment benefit in order to guide personalized therapeutic decisions. AutoPrognosis is an (open-source) software (https://bitbucket.org/mvdschaar/mlforhealthlabpub) that we have developed to automate the deployment of machine learning in clinical prognostic modeling. The AutoPrognosis algorithm automatically generates a bespoke machine learning model for the data set at hand by optimizing an ensemble of machine learning models (e.g., neural networks, random forests, etc.) using an advanced Bayesian optimization algorithm, and then uses a symbolic regression algorithm[20]

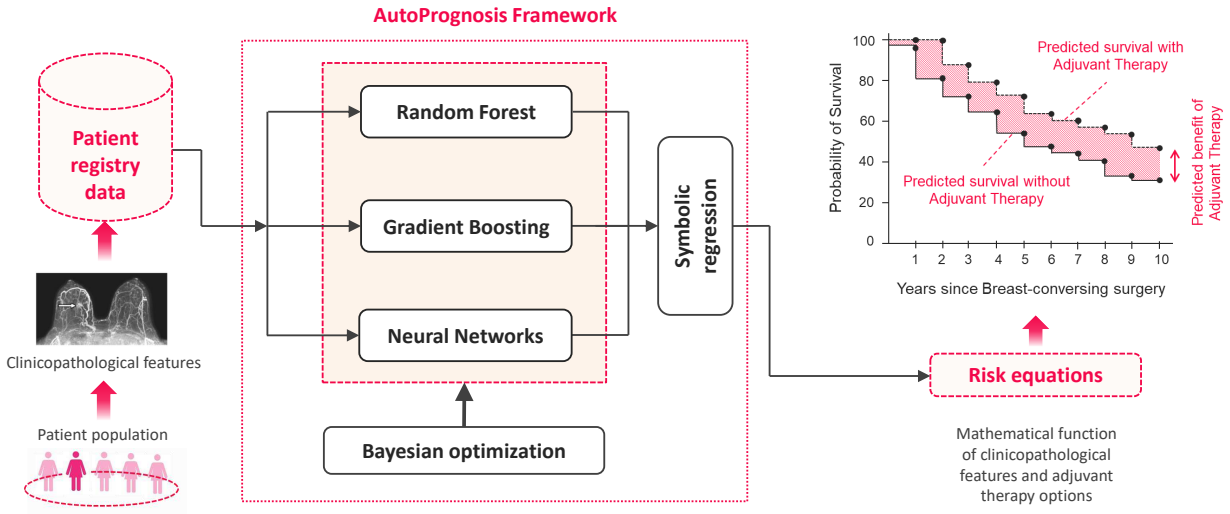


Figure 1: **Schematic depiction of the AutoPrognosis framework.** Given patient data, AutoPrognosis uses a Bayesian optimization algorithm to search for the optimal parameters of a collection of machine learning models and the optimal weight assigned to each model in an ensemble. (Here, we depict random forests, gradient boosting and neural network models as exemplary elements of the ensemble.) After fitting the ensemble model, a symbolic regression algorithm is used to convert the fitted model into a mathematical equation that maps patient variables to predicted risk. The end result is a mathematical equation that computes an individual patient's survival curve with and without a given therapy.

to convert the optimized ensemble into a transparent risk equation that is interpretable to clinicians (Fig. 1). We developed and validated Adjutorium through the AutoPrognosis software using data for nearly 1 million women in large-scale cohorts that are representative of the UK and US populations.

We trained Adjutorium to predict breast cancer and all-cause mortality without adjuvant therapies by fitting 10 binary classification ensemble models (optimized via AutoPrognosis), where each model was trained to predict patient survival at 10 distinct time horizons spanning from 1 to 10 years from baseline, with 1-year increments. The effects of four adjuvant therapies (chemotherapy, hormone therapy, bisphosphonates and trastuzumab) were incorporated into the model using their estimated relative risk reduction rates from the EBCTCG meta-analysis.[21, 22] The input to the model is a set of features for an individual patient, and the outputs are the patient's predicted (breast cancer-specific and all-cause) survival curves under no adjuvant therapy and any combination of the four adjuvant therapies under consideration (inputs and outputs for Adjutorium are visualized in the following web application: https://adjutorium-breastcancer.herokuapp.com/). Technical details for the implementation of AutoPrognosis have been described previously.[20, 23, 24] A brief discussion of AutoPrognosis and a detailed explanation of the training procedure for Adjutorium are provided in Methods and Supplementary Information.

Through internal and external validation, we compared the accuracy of Adjutorium in predicting all-cause and breast cancer-specific mortality at 3, 5 and 10 years from baseline with the commonly used PREDICT v2.1 score,[7] in addition to an in-house Cox proportional hazards (PH) regression model fitted to the same training cohort used to derive the Adjutorium model. We assessed the discriminative accuracy of all models using the time-dependent area under receiver operating characteristic curve[25] (AUC-ROC), Harrell's concordance index[26] (C-index), and Uno's C-index.[27] Details on the mathematical definitions of each of these metrics can be found in Supplementary Information. For all evaluations, 95% confidence intervals on the estimated performance metrics were obtained via bootstrapped re-sampling of the validation data.

Data resources and study cohorts

Patient data for the study were obtained from two cohorts: the UK National Cancer Registration and Analysis Service (NCRAS, n=620,249), and the US Surveillance, Epidemiology and End Results program[28] (SEER, n=588,735). NCRAS is the population-based cancer registry for England; the SEER program at the National Cancer Institute collects data on cancer diagnoses, treatment and survival for approximately 30% of the US population. The two databases combined hold data for over 1.2 million cases diagnosed between 2000 and 2016. Data was extracted for early breast cancer patients — patients with metastatic cancer were excluded. We extracted patient-level data: patients with multiple primary tumors were represented through their first diagnosis only. The extracted patient-level data comprised standard prognostic factors used in existing prognostic models,[7, 29, 30] including age at diagnosis, mode of detection (screen-detected/symptomatic), estrogen receptor (ER) status, human epidermal growth factor receptor 2 (HER2) status, number of lymph nodes involved, tumour size and histological tumour grade. As this was a large population-based study, with full anonymisation of all data, informed consent and ethical approval was not sought.

A total of 395,862 and 571,635 patients met the inclusion criteria in NCRAS and SEER, respectively (Supplementary Fig. 1). Missing data was imputed using the multiple chained equations[31] (MICE) method. Details on the patient inclusion criteria and the steps involved in missing data imputation are provided in Methods and Supplementary Information; patient characteristics are provided in Supplementary Table 1. Patient samples from the NCRAS database were randomly split into two mutually exclusive cohorts: a training cohort of 316,690 patients used for model derivation, and an internal validation cohort of 79,172 patients used to evaluate model accuracy. The entire SEER cohort (571,635 patients) was reserved for external validation. The primary outcome of our study

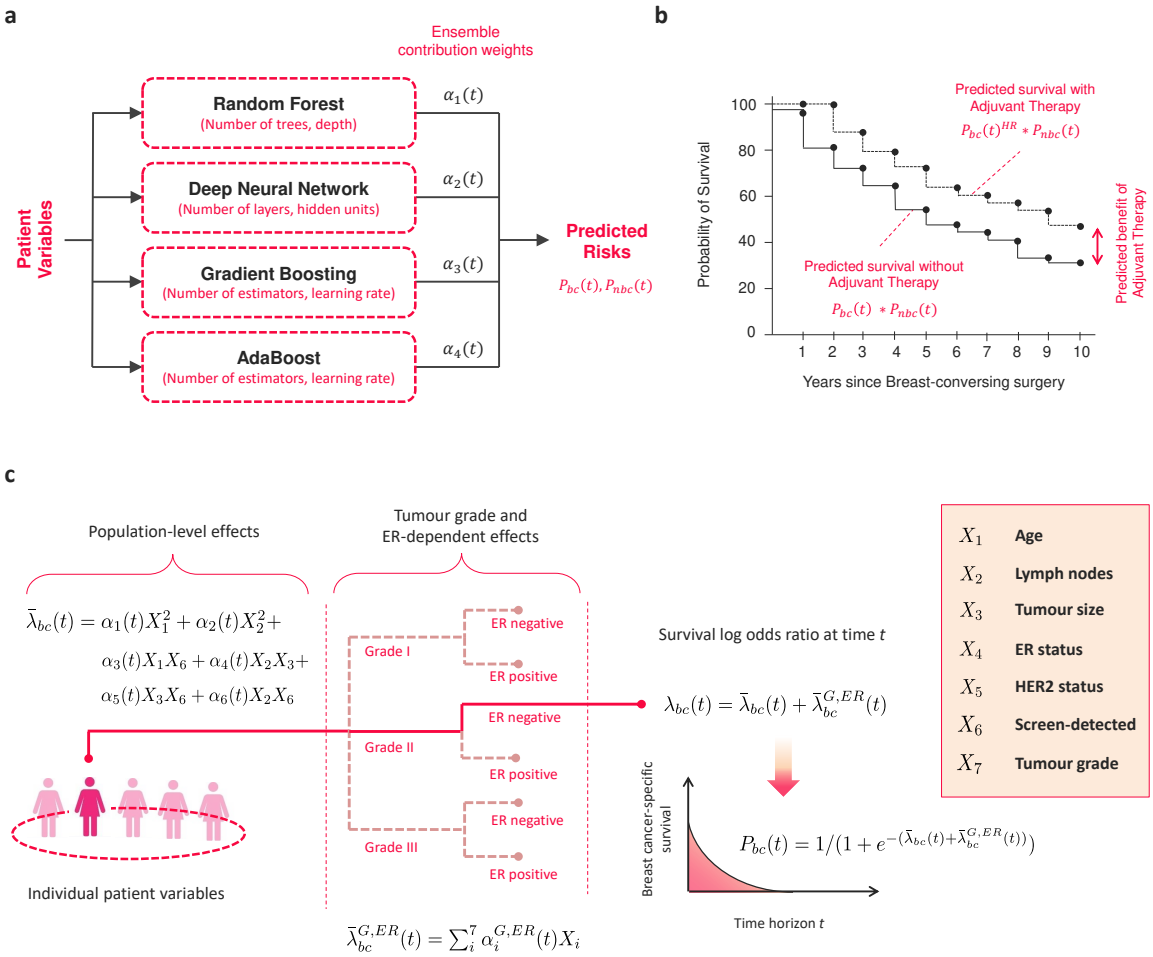


Figure 2: **Illustration for the machine learning model underlying Adjutorium. a,** *The ensemble model learned by the AutoPrognosis software.* The ensemble comprises four basic machine learning models: random forest, neural network, gradient boosting, and AdaBoost. The prediction issued by Adjutorium is a weighted combination of the predictions of the four members of the ensemble. Each model in the ensemble has a set of parameters (listed between brackets), and an assigned weight $\alpha(t)$ determining its contribution in the final prediction. Both the model parameters and its weight change depending on the prediction horizon t. Separate ensembles are trained to predict breast cancer-specific survival $P_{bc}(t)$ and other cause survival $P_{nbc}(t)$. **b,** *The predicted survival curve for an exemplary patient (with and without adjuvant therapy).* Here, each prediction horizon (1 to 10 years since diagnosis, with 1-year steps) corresponds to a knot in the survival curve, and each knot is associated with a distinct set of model parameters and contribution weights in the ensemble in **a**. **c,** Risk equations underlying Adjutorium as learned by the symbolic regression module in AutoPrognosis. Given the individual-level variables of a patient, the risk equation evaluates the probability of survival at future time horizons. The log odds ratio for survival at time t comprises two components: (1) a population-level term that models non-linear effects of age and number of lymph nodes, in addition to interactions between different variables through six coefficients that are fixed for all patients, and (2) a tumour grade and ER-specific term that evaluates the linear effects of all prognostic factors with coefficients that are specific to every group of patients with the same grade and ER status. Here we show an exemplary patient with ER negative cancer and tumour grade 2 and. The risk equation is a mathematical abstraction for the predictions issued by the machine learning model in **a**.

was survival from all-cause mortality at 3, 5 and 10 years after surgery for breast cancer. All-cause mortality was further subdivided into breast cancer-specific mortality, which was assessed as a secondary outcome, and mortality due to other causes. Breast cancer-specific mortality was defined as ICD-10 code C.50 listed on the death certificate as a cause of death.

Development of the Adjutorium model for breast cancer prognostication

A high-level illustration for the machine learning model generated by AutoPrognosis when fitted to the development cohort (n=316,690) is provided in Fig. 2. The overall model is based on two ensembles, each comprising four binary classification models:[32] random forest, neural network, gradient boosting, and AdaBoost. One ensemble was trained to predict the risk of breast cancer-specific mortality $P_{bc}(t)$ at a time horizon t based on all prognostic variables, and the other ensemble was trained to predict the risk of other cause mortality $P_{nbc}(t)$ based on age. All-cause survival was computed as $P_{bc}^{HR}(t) \cdot P_{nbc}(t)$, where HR is the risk reduction rate ratio (hazard ratio) of the selected adjuvant therapy ($HR = 1$ if no treatment is administered). The values of HR for chemotherapy, hormone therapy, bisphosphonates and trastuzumab were obtained from the EBCTCG meta-analyses.[21, 22]

Through the symbolic regression module in AutoPrognosis (Fig. 1), the ensemble model for $P_{bc}(t)$ was mathematically represented in the form of a risk equation that maps patient variables to breast-cancer-specific survival functions (See Fig. 2(c) for a visual depiction of this equation). The risk equation for $P_{bc}(t)$ can be described as follows. For a given patient, breast-cancer-related survival probability is given by $P_{bc}(t) = 1/(1+\exp(-\lambda_{bc}(t)))$, where t is the time horizon at which the survival probability is evaluated. The term $\lambda_{bc}(t)$ can be interpreted as the *log odds ratio* for survival at time t, and it comprises the following two components:

$$\lambda_{bc}(t) = \underbrace{\bar{\lambda}_{bc}(t)}_{\text{Population-level}} + \underbrace{\bar{\lambda}_{bc}^{G,ER}(t)}_{\text{Grade-ER-specific}},$$

where the first term $\bar{\lambda}_{bc}(t)$ comprises coefficients shared among all patients in the population, and includes the non-linear effects of the age and number of lymph nodes variables, in addition to interaction terms between age, mode of detection, tumour size and number of lymph nodes (Fig. 2(c)). These interaction terms reflect the impact of the implemented screening policy on patients' risks, i.e., the coefficients $(\alpha_3, \alpha_5, \alpha_6)$ in Fig. 2(c) quantify the risk reduction (by early detection of cancer via screening) as a function of the patient's age and tumor spread at diagnosis time. The second term, $\bar{\lambda}_{bc}^{G,ER}(t)$, includes linear contributions of all prognostic variables, with coefficients

specific to subgroups of patients with every possible combination of tumour grade and ER status. The numerical values of the coefficients of $\lambda_{bc}(t)$ are provided in the Supplementary Information.

The breast cancer-specific mortality risk equation learned by AutoPrognosis demonstrates that our machine learning approach identified new interactions that were not incorporated in previous models[7], namely the interactions between tumour grade and all other variables. These results are in agreement with new approaches to molecular subtyping that use both receptor status and tumor grade to categorize breast cancer into several conceptual molecular classes (e.g., Luminal A and B subtypes) that have different prognoses and (potentially) different responses to specific therapies.[33] Thus, the interpretable risk equation learned by AutoPrognosis not only ensures model transparency, but also provides insights into the discovery of new breast cancer subtypes.

For benchmark purposes, the PREDICT v2.1 score and a standard Cox PH model fit on the same training data as Adjutorium were also assessed for comparison. Consistent with previous studies,[7] we fitted two separate Cox models, with different baseline hazards for ER positive and ER negative cancer to capture the interactions between ER status and other prognostic variables. We included an age squared term to allow for non-linear effects of baseline age at diagnosis on breast cancer mortality. Tumor size and number of lymph nodes were both coded as continuous variables. Coefficients of the fitted Cox PH model are provided in Supplementary Table 2.

Accuracy of the Adjutorium model

Of 395,862 eligible patients in NCRAS, the mean age of breast cancer diagnosis was 61 years, with 2 million person-years of total follow-up (median follow-up time of 5.2 years) within the cohort. The SEER cohort included 571,635 eligible patients with a mean age of diagnosis of 61 years, and a total 3.2 million person-years of follow-up (median follow-up time of 5.7 years). During follow-up, 83,139 and 139,225 deaths were recorded in NCRAS and SEER, respectively, of which 53,143 (64%) and 59,585 (43%) cases were breast cancer-related. Overall 5-year survival from breast cancer were 90% and 86% in SEER and NCRAS, respectively.

Discriminative accuracy. Adjutorium uniformly outperformed PREDICT v2.1 and the conventional Cox PH model in predicting all-cause and breast cancer-specific mortality, both when validated internally within NCRAS, and externally within the SEER cohort (Table 1). The improvements were achieved with respect to all discriminative accuracy metrics and all time horizons under study.

	Time Horizon	Metric (95% CI)	All-cause Mortality			Breast cancer-specific Mortality		
			Adjutorium	Cox PH	PREDICT	Adjutorium	Cox PH	PREDICT
Internal Validation Cohort (NCRAS, n=79,172)	3 years	H. C-index	0.782 (0.781–0.783)	0.755 (0.753–0.757)	0.746 (0.745–0.747)	0.809 (0.808–0.810)	0.773 (0.771–0.775)	0.739 (0.738–0.740)
		U. C-index	0.755 (0.753–0.757)	0.735 (0.733–0.737)	0.708 (0.705–0.711)	0.764 (0.762–0.766)	0.732 (0.730–0.734)	0.701 (0.700–0.702)
		AUC-ROC	0.818 (0.816–0.820)	0.795 (0.793–0.797)	0.785 (0.783–0.787)	0.849 (0.847–0.851)	0.817 (0.816–0.818)	0.766 (0.764–0.768)
	5 years	H. C-index	0.787 (0.785–0.789)	0.755 (0.753–0.757)	0.757 (0.755–0.759)	0.808 (0.807–0.809)	0.774 (0.773–0.775)	0.749 (0.748–0.750)
		U. C-index	0.755 (0.753–0.757)	0.733 (0.732–0.734)	0.718 (0.716–0.720)	0.767 (0.765–0.769)	0.737 (0.735–0.739)	0.709 (0.707–0.711)
		AUC-ROC	0.816 (0.814–0.818)	0.773 (0.771–0.775)	0.775 (0.773–0.777)	0.835 (0.833–0.837)	0.796 (0.794–0.798)	0.755 (0.753–0.757)
	10 years	H. C-index	0.773 (0.771–0.775)	0.759 (0.757–0.760)	0.772 (0.770–0.774)	0.790 (0.788–0.792)	0.778 (0.777–0.779)	0.751 (0.749–0.753)
		U. C-index	0.745 (0.743–0.747)	0.735 (0.734–0.736)	0.734 (0.732–0.736)	0.756 (0.754–0.758)	0.736 (0.734–0.738)	0.715 (0.714–0.716)
		AUC-ROC	0.815 (0.813–0.817)	0.775 (0.773–0.777)	0.770 (0.768–0.772)	0.825 (0.823–0.827)	0.783 (0.781–0.785)	0.730 (0.727–0.733)
	Time Horizon	Metric (95% CI)	All-cause Mortality			Breast cancer-specific Mortality		
			Adjutorium	Cox PH	PREDICT	Adjutorium	Cox PH	PREDICT
External Validation Cohort (SEER, n=571,635)	3 years	H. C-index	0.752 (0.749–0.755)	0.746 (0.745–0.747)	0.737 (0.736–0.738)	0.797 (0.795–0.799)	0.763 (0.760–0.766)	0.764 (0.762–0.766)
		U. C-index	0.743 (0.741–0.745)	0.735 (0.734–0.736)	0.698 (0.696–0.700)	0.755 (0.750–0.760)	0.727 (0.722–0.732)	0.721 (0.715–0.727)
		AUC-ROC	0.771 (0.770–0.772)	0.773 (0.770–0.776)	0.762 (0.761–0.763)	0.823 (0.820–0.826)	0.792 (0.787–0.797)	0.784 (0.782–0.786)
	5 years	H. C-index	0.758 (0.757–0.759)	0.744 (0.742–0.746)	0.743 (0.741–0.745)	0.796 (0.794–0.798)	0.769 (0.766–0.772)	0.765 (0.763–0.767)
		U. C-index	0.736 (0.732–0.740)	0.732 (0.725–0.739)	0.709 (0.707–0.711)	0.760 (0.755–0.765)	0.722 (0.714–0.730)	0.735 (0.730–0.740)
		AUC-ROC	0.777 (0.775–0.779)	0.763 (0.759–0.767)	0.755 (0.753–0.757)	0.815 (0.813–0.817)	0.784 (0.782–0.786)	0.775 (0.772–0.778)
	10 years	H. C-index	0.749 (0.746–0.752)	0.741 (0.737–0.745)	0.751 (0.750–0.752)	0.778 (0.776–0.780)	0.764 (0.761–0.767)	0.765 (0.763–0.767)
		U. C-index	0.735 (0.730–0.740)	0.738 (0.732–0.744)	0.728 (0.726–0.730)	0.746 (0.741–0.751)	0.728 (0.720–0.736)	0.738 (0.734–0.742)
		AUC-ROC	0.790 (0.787–0.793)	0.778 (0.771–0.785)	0.756 (0.753–0.759)	0.803 (0.800–0.806)	0.775 (0.770–0.780)	0.744 (0.741–0.747)

* CI denotes Confidence Interval. H. C-index and U. C-index denote the Harrell and Uno concordance indexes, respectively.

Table 1: Discriminative accuracy with respect to the primary and secondary outcomes.

In internal validation, Adjutorium predicted 10-year all-cause mortality with an AUC-ROC accuracy of 0.815 (95% CI: 0.813-0.817), compared with 0.777 (95% CI: 0.768-0.772) by PREDICT v2.1, and 0.775 (95% CI: 0.773-0.777) by the Cox PH model. Similar performance gains were achieved over the other time horizons, and with respect to the C-index statistic (Table 1). The improvements in accuracy achieved by Adjutorium were even more significant in predicting breast cancer-specific mortality, with an AUC-ROC of 0.825 (95% CI: 0.823-0.827) for 10-year outcomes, compared with 0.730 (95% CI: 0.727-0.733) by PREDICT v2.1, and 0.783 (95% CI: 0.781-0.785) by the Cox PH model. The fact that the accuracy improvements were more significant in the secondary outcome is not surprising since all of the variables included in the model were breast cancer-related.

Adjutorium generalized well to the external validation cohort, with similar accuracy improvements for both the primary and secondary outcomes (**Table 4**). With respect to 10-year all-cause mortality, Adjutorium achieved an AUC-ROC of 0.790 (95% CI: 0.787-0.793), compared to 0.756 (95% CI: 0.753-0.759) by PREDICT, 0.631 (95% CI: 0.628-0.634) by NPI, and 0.778 (95% CI: 0.771-0.785) by the Cox PH model. Similar gains were achieved over the other time horizons (**Table 4**). For prediction of 10-year breast cancer-specific mortality, Adjutorium achieved an AUC-ROC of 0.803 (95% CI: 0.800-0.806), compared to 0.744 (95% CI: 0.741-0.747) by PREDICT, 0.768 (95% CI: 0.765-0.771) by NPI, and 0.775 (95% CI: 0.770-0.780) by the Cox PH model.

Importantly, Adjutorium outperformed the Cox PH model fitted to the same development cohort, reflecting the *gain from modeling*, i.e., the gain achieved by using flexible machine learning models instead of standard regression. On the other hand, the gain achieved by the Cox PH model compared to PREDICT v2.1 in external validation reflects the *gain from information*, i.e., the gain achieved by using large-scale, representative data that enhance the accuracy and generalizability of the fitted models to other cohorts that might entail different demographic structure and outcomes.

Subgroup analysis. The accuracy improvements achieved by Adjutorium were consistent across all subgroups of patients stratified by age, HER2 status, ER status and tumour grade (Table 2). Improvements were greater in subgroups that are poorly served by current prognostic tools; the accuracy gains achieved by Adjutorium relative to PREDICT v2.1 were higher in elderly patients (age > 65 yrs at diagnosis), patients with ER negative and HER2 negative breast cancer. This is likely due to the fact that our machine learning-based risk equation captured nuanced interactions and non-linear patterns that were not incorporated in existing prognostic tools (Fig. 2(c)).

				Adjutorium			PREDICT v2.1	
	No. of cases	Observed deaths	AUC-ROC	TP	FP	AUC-ROC	TP	FP
Internal validation cohort (NCRAS)								
Age at diagnosis								
ER positive								
30 – 65 years	21,302	2,314	0.791	1,658	5,142	0.773	1,607	5,171
> 65 years	13,115	3,774	0.824	3,026	2,767	0.779	2,915	2,937
ER negative								
30 – 65 years	10,417	2,440	0.729	1,615	2,634	0.666	1,595	3,043
> 65 years	4,861	2,090	0.785	1,458	730	0.700	1,626	1,202
HER2 positive								
30 – 65 years	11,894	2,390	0.717	1,563	3,157	0.682	1,535	3,299
> 65 years	4,388	1,940	0.767	1,370	733	0.671	1,449	1,131
HER2 negative								
30 – 65 years	19,825	2,363	0.816	1,749	4,286	0.797	1,749	4,898
> 65 years	13,588	3,924	0.825	2,970	2,443	0.763	3,088	3,433
Grade I								
30 – 65 years	4,942	146	0.752	101	1,262	0.739	103	1,580
> 65 years	2,608	382	0.816	273	423	0.758	290	683
Grade II								
30 – 65 years	6,472	1,772	0.753	1,218	1,286	0.720	1,291	1,754
> 65 years	6,920	2,891	0.693	2,120	1,737	0.684	2,165	1,824
Grade III								
30 – 65 years	5,935	2,820	0.730	2,061	1,210	0.630	1,785	1,249
> 65 years	4,503	2,577	0.662	1,921	942	0.613	1,370	652
				Adjutorium			**PREDICT v2.1**	
	No. of cases	Observed deaths	AUC-ROC	TP	FP	AUC-ROC	TP	FP
External validation cohort (SEER)								
Age at diagnosis								
ER positive								
30 – 65 years	74,732	18,374	0.798	14,286	20,034	0.799	14,941	19,544
> 65 years	38,226	14,290	0.806	10,527	6,174	0.800	10,688	6,212
ER negative								
30 – 65 years	61,070	17,594	0.768	11,552	11,138	0.727	10,829	12,948
> 65 years	25,812	11,564	0.797	7,894	3,378	0.766	9,053	4,571
HER2 positive								
30 – 65 years	1,467	1,467	___	___	___	___	___	___
> 65 years	958	958	___	___	___	___	___	___
HER2 negative								
30 – 65 years	134,335	34,501	0.766	24,155	33,485	0.745	23,441	27,704
> 65 years	63,080	24,896	0.791	17,383	9,978	0.769	16,206	8,243
Grade I								
30 – 65 years	18,073	1,517	0.736	1,025	4,139	0.725	960	5,575
> 65 years	10,643	1,850	0.740	1,110	2,146	0.700	983	1,849
Grade II								
30 – 65 years	68,596	13,180	0.718	9,691	12,477	0.712	6,736	11,387
> 65 years	63,009	13,397	0.700	9,433	21,828	0.700	7,584	12,329
Grade III								
30 – 65 years	62,560	21,157	0.728	13,754	12,029	0.730	13,456	11,885
> 65 years	30,630	10,531	0.700	6,360	7,829	0.720	7,724	7,820

* FP and TP denote false positive and true positive cases, respectively.

Table 2: Subgroup-level discrimination with respect to breast cancer-specific 10-year outcomes.

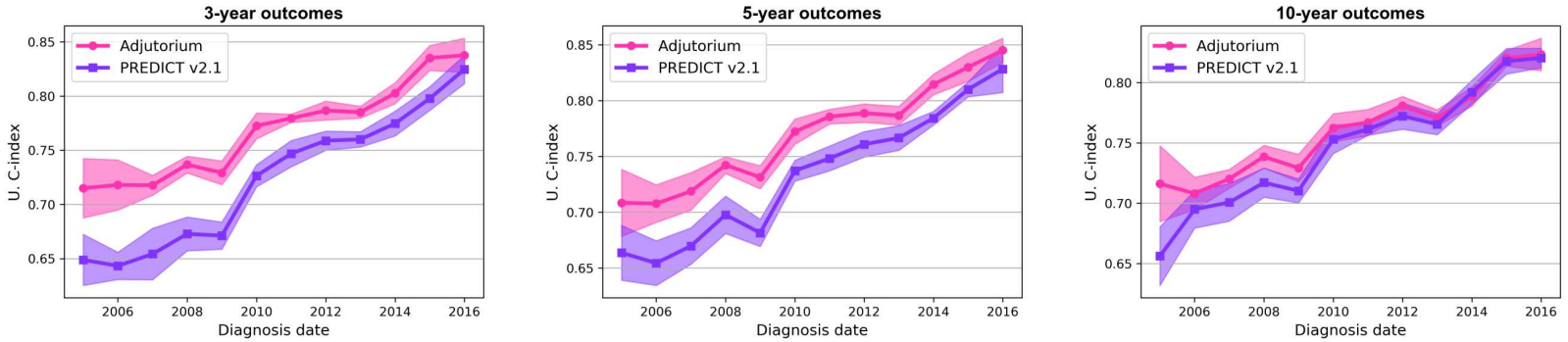
Sensitivity analyses and calibration performance. We conducted various tests to evaluate the robustness of our results. First, we tested the robustness of Adjutorium to time-cohort effects; internal validation on sub-cohorts stratified by diagnosis dates from 2005 to 2016 showed that the accuracy gains by Adjutorium are achieved for all diagnosis years, except for 10-year all-cause mortality in more recent diagnosis years where both models perform similarly (Fig. 3). (This is mainly because recent cohorts do not have sufficient follow-up.) Moreover, we applied internal and external validation on sub-cohorts with complete data and missing data to test the robustness of Adjutorium to data missingness; the model performed well in cases with complete and missing data, outperforming other models by similar margins in both analyses (Supplementary Information). When validated on 21,164 patients (in the internal validation cohort) with complete data on all variables, the AUC-ROC accuracy of Adjutorium with respect to 10-year breast cancer-specific mortality was 0.811 (95% CI: 0.0.808-0.814), and 0.783 (95% CI: 0.780-0.786) for PREDICT v2.1. When validated on 57,996 patients with missing data on one or more variables, the AUC-ROC accuracy of Adjutorium was 0.829 (95% CI: 0.0.827-0.831), and 0.728 (95% CI: 0.725-0.731) for PREDICT v2.1.

Adjutorium was well-calibrated across study cohorts, displaying better calibration with observed outcomes than PREDICT v2.1 (Supplementary Information). In internal validation, we found that PREDICT v2.1 substantially over-estimated the risk of both all-cause and breast cancer related mortality at 10-year follow up. In external validation, PREDICT v2.1 over-estimated the risk of breast cancer related mortality, but was relatively more conservative in predicting all-cause mortality. While Adjutorium was noted to under-estimate mortality in patients who were at high risk for breast cancer and all cause mortality, this is unlikely to impact clinical decision making as these individuals are likely to be well beyond the decision threshold for improvement with treatment. Moreover, patients in this risk subgroup comprised only 6% of the overall population.

Impact on adjuvant therapy decisions

To assess the clinical benefit of using Adjutorium for supporting decisions regarding adjuvant therapies, we compared Adjutorium predictions of treatment benefit to PREDICT v2.1, and the observed decisions of multidisciplinary teams (MDT) obtained from the NCRAS database. To this end, we followed decision thresholds currently used for decision-making with PREDICT within the UK, recommending chemotherapy if a patient's 10-year net survival benefit from treatment is predicted to be greater than 5%[34] and no adjuvant chemotherapy if treatment benefit is <3%. The decisions when survival benefit is predicted as 3-5% are made on a case by case basis, and no

a Discriminative accuracy with respect to all-cause mortality



b Discriminative accuracy with respect to breast cancer-specific mortality

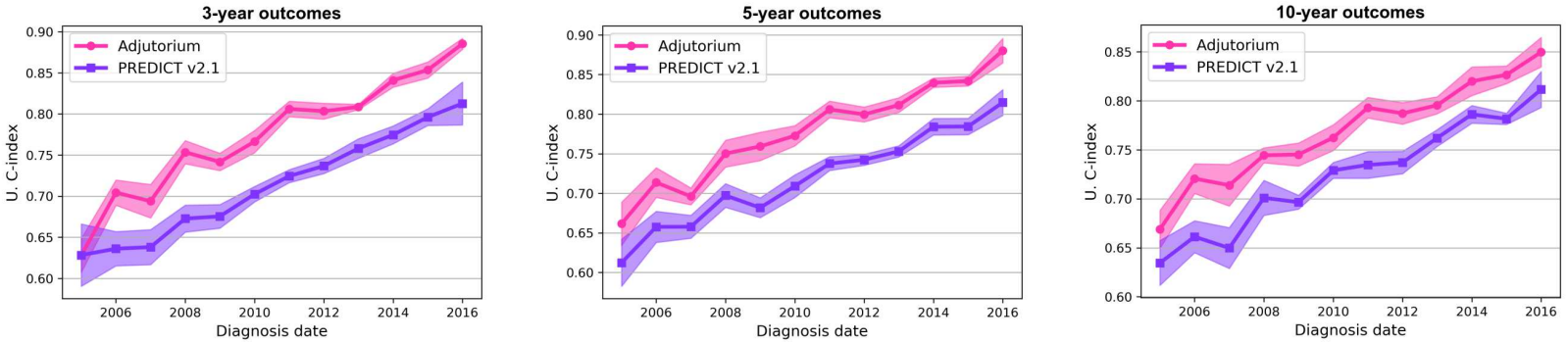


Figure 3: Discriminative accuracy evaluated in sub-cohorts of patients stratified by diagnosis date.

formal guidelines exist regarding these at present. We compared 5- and 10-year survival among patients where MDT decision-making regarding treatment (extracted from the registry data) had been concordant with Adjutorium, with survival among patients where this had been discordant. We also conducted a similar comparison with PREDICT v2.1 examining average survival of patients with discordant predictions of treatment benefit between the algorithms. Finally, we assessed how many additional patients who had died of breast cancer within 10 years would have been assigned to treatment by Adjutorium relative to treatment assignment by MDTs and PREDICT v2.1.

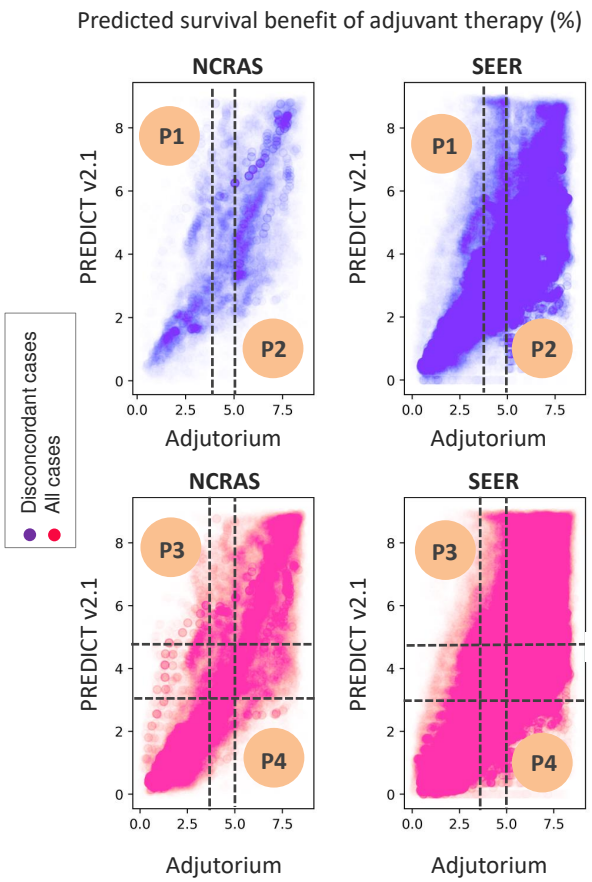
The average benefit of chemotherapy predicted by Adjutorium and PREDICT v2.1 in all study cohorts were found to be significantly different (t-test, $p < 0.001$). Fig. 4 visualizes the disconcordance between treatment decisions informed by Adjutorium and PREDICT v2.1, in addition to the observed MDT decisions in light of the patients' 10-year outcomes. In both the internal and external validation cohorts, Adjutorium and PREDICT v2.1 disagreed on treatment decisions for 19% of the patient population (Fig. 4(a)). The population of patients that were recommended a treatment by Adjutorium but not by PREDICT or MDTs (Population P2 and P4 in Fig. 4) had a higher than average mortality rate at 10 years. An average 10-year mortality of 28% is consistent with a benefit of >5%, suggesting that on average, this treatment subgroup would have benefited from treatment.

On the contrary, the population of patients that were not recommended a treatment by Adjutorium, but were recommended a chemotherapy by PREDICT or the MDT decisions exhibited a 10-year mortality rate less than that of the populational average. A 10-year mortality of 18% in the group discordantly assigned to treatment by PREDICT suggests average treatment benefit in the range of 2.4%. This indicates that treatment decisions informed by Adjutorium are less likely to over or under treat patients. Compared to historical decisions made by multidisciplinary teams (MDT), Adjutorium can potentially improve treatment decisions for 25% of the patient population (13% who are under-treated, and 12% who are potentially over-treated).

Discussion

We developed and validated Adjutorium — a machine learning-based tool for predicting the individualized benefit of adjuvant therapies in breast cancer. Involving data from nearly 1 million individuals with breast cancer from the UK and US, this is one of the largest studies of its kind. We found that Adjutorium substantially outperforms one of the most widely used standards for clinical decision making, and critically is generalisable to distinct clinical settings. To our knowledge this is

a Disconcordance between different models



b Characteristics of patients in disconcordant cases

Population P1: untreated by Adjutorium, treated by MDT. **MR*:** NCRAS=0.85, SEER=0.88.

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=9,904)	55.7	2.5	22	10%	69%	41%	65%
SEER (n=73,866)	57.7	1.9	22	18%	80%	13%	___

Population P2: treated by Adjutorium, untreated by MDT. **MR:** NCRAS=1.13, SEER=1.08.

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=10,345)	63.0	5.6	28	2.7%	59%	37%	16.7%
SEER (n=101,861)	57.3	1.2	24	1.8%	41%	3%	___

Population P3: untreated by Adjutorium, treated by PREDICT. **MR:** NCRAS=0.92, SEER=0.75.

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=5,096)	60.0	7.9	23	6%	43%	47%	34%
SEER (n=12,105)	60.0	5.9	37	2%	29%	15%	___

Population P4: treated by Adjutorium, untreated by PREDICT. **MR:** NCRAS=1.20, SEER=1.60

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=9,285)	58.2	1.9	25	1.9%	66%	20%	41%
SEER (n=99,389)	55.4	0.7	20	2.5%	54%	3.5%	___

* MR stands for “mortality ratio”, defined as the ratio between the 10-year mortality rate in the selected population and that of the overall population.

Figure 4: Comparison between therapeutic decisions informed by Adjutorium and PREDICT v2.1.

the first application of a machine learning model for prognostication in breast cancer, that has been shown to be generalisable across multiple nationally representative cohorts.

While several prognostication methods are available for supporting clinical decisions regarding adjuvant therapies in breast cancer, they have well recognized limitations particularly in terms of their accuracy in certain subgroups and their generalisability to other populations. We find that Adjutorium outperforms existing clinical decision support tools in terms of accuracy, and calibration to observed outcomes, across all patient groups. Additionally, it shows substantially improved performance in subgroups where existing clinical decision support tools are known perform poorly (e.g., older women with early cancer, and ER negative breast cancer) suggesting that using Adjutorium to support clinical decisions may lead to better treatment decisions, and potentially better outcomes in these subgroups. By contrast with other existing tools, Adjutorium is robust to missing data, and is able to make accurate predictions even when information on some of the prognostic factors is not available. This is an important advance, making our model more generalisable to settings where data on patients may be incomplete. Importantly, we observe lower 10-year mortality among patients where MDT decisions are concordant with Adjutorium predictions; this has important implications for clinical decision support, and highlights the utility of tools such as Adjutorium for prognostication to potentially drive better patient outcomes.

We find that Adjutorium not only outperforms PREDICT v2.1, but also a Cox proportional hazards model fit on the same training cohort. This suggests that gains in performance are achieved not only due to a larger representative set for training the models, but also due to the flexible nature of the machine learning algorithms applied. Our fitted model does not make any assumptions about the linearity of the patient risks as function of prognostic factors, or the proportionality of hazards over time. Additionally it is able to infer interactions, and non-linear associations in a data-driven fashion, as evident through the interpretable risk equations describing the machine learning model.

In order to improve accessibility and general use, we also provide an easy-to-use online tool for breast cancer prediction (http://www.vanderschaar-lab.com/adjutorium/) based on the Adjutorium model, where patient feature can be easily input to a visualization tool that depicts the patient survival time under different treatment options. This portal allows clinicians to work with patients to make important decisions regarding adjuvant therapy treatments in a personalized context. We, therefore, provide an important clinical tool for breast cancer treatment management to be used within the UK, and globally. Moreover, we provide an open-source software for the AutoPrognosis

system, which enables other researchers to easily re-fit the model as more data becomes available. Because our approach is automated, it would help clinical researchers update the model as different aspects of the health care system change over time (e.g., introduction of novel adjuvant therapies), without the need to involve experts in making new modeling choices and decisions repeatedly for every new update. Moreover, the symbolic regression module can communicate these model updates with clinicians by highlighting changes in model coefficients and newly discovered interactions and non-linearity, which makes the entire process transparent.

We acknowledge limitations of our model, which include the retrospective nature of our study which makes it difficult to assess changes in patient outcomes when using Adjutorium relative to existing tools. Another limitation is that our model does not predict outcomes such as recurrence, and currently does not incorporate multigene assays or other gene expression-based predictive information. However, these can be easily incorporated into our model. Also, Adjutorium does not explicitly derive treatment effects in a data-driven fashion, rather using estimates from meta-analyses on clinical trials. We also acknowledge limitations of the data used to derive our model, which include the lack of complete information on bisphosphonates and trastuzumab in the NCRAS derivation cohort, lack of information on treatments other than chemotherapy in SEER and incomplete coding of chemotherapy variables in SEER. Using our automated algorithm, the model can be easily updated once complete information on these treatments become available.

In summary, we have developed and validated Adjutorium, a flexible and generalizable machine learning-based tool for clinical decision support in breast cancer treatment. Our work suggests that using Adjutorium to support decisions made by multidisciplinary teams around adjuvant therapy could potentially improve patient outcomes relative to existing decision support tools, across distinct clinical settings. Further work in prosective longitudinal cohort studies will be needed to quantify and realise these benefits in practice.

Methods

Data sources and patient inclusion criteria

From NCRAS, we included patients who were diagnosed after January 1^{st} 2005. This extra inclusion criteria in NCRAS was necessary as the missingness of HER2 status variable was predictive of the outcome (i.e., patients who have HER2 missing has worse outcomes on average). Because the missingness rate of HER2 prior to 2005 was very high, including patients with complete HER2 information who were diagnoses dates prior to 2005 would cause a bias in the survival outcomes. From both datasets, we included patients who were aged 30 to 90 years at diagnosis. Specific age data were not available on patients less than 30 years of age in NCRAS; hence, these were excluded. Furthermore, we excluded patients with missing data on more than 4 variables (<10$\%$ of all participants), and a small number of patients who were outliers for tumour size (>90 mm tumour), and number of positive lymph nodes (>50). A total of 395,862 and 571,635 patients met the inclusion criteria in NCRAS and SEER, respectively. We did not include Ki67 as it was not available for the vast majority of patients in NCRAS, and has already been shown to have poor predictive power.[35, 36]

The extracted NCRAS dataset contained complete information on which patients where treated with chemotherapy and hormone therapy, but did not include information on other adjuvant therapies, such as targeted anti-HER2 agents. Release of complete treatment information was in violation of the data anonymisation constraints imposed by the NCRAS data sharing policy; in addition, information on other adjuvant therapies was only routinely recorded for patients diagnosed in more recent years. Thus, to validate our model on data with complete treatment information, we acquired an anonymised supplementary NCRAS dataset of 17,804 patients diagnosed in 2013, with complete information on chemotherapy, hormone therapy, immunotherapy, CDK4/6 inhibitors, PARP inhibitors, Trastuzumab and Bisphosphonates. We denote this dataset as NCRAS-2; details on the patient characteristics and validation results on the NCRAS-2 sub-cohort is provided in Supplementary Information. The NCRAS-2 sub-cohort (including all patients diagnosed in 2013 within the NCRAS data set) comprised a total of 17,804 eligible patients with a median follow-up time of 5.38 years. Among these, 84.72$\%$ received chemotherapy, 19.49$\%$ received hormone therapy, 22.43$\%$ received trastuzumab and 3$\%$ received bisphosphonates.

Missing data imputation

A limitation of existing models has been their dependence on complete case analysis, and lack of flexibility to incorporate missing variables. Our analysis suggested that missingness was highly informative;[37] (log-rank test for difference in 5-year survival between patients with complete data and one or more missing variable, $p < 0.001$). In this context, including only patients with complete data is likely to affect model generalisability. Therefore, in the interest of generalisability, we opted to impute any missing data using data

available on other variables. For all study cohorts, we imputed missing data using the model-based multiple chained equations[31] (MICE) method. We create 10 imputed datasets and pool the predictions of all models under study using Rubin's rule.[38] Details regarding imputation are provided in Supplementary Information.

Model development

Automated machine learning. We derived the Adjutorium model using the *AutoPrognosis*[19] framework, an (open-source) software (https://bitbucket.org/mvdschaar/mlforhealthlabpub) that we have developed to automate the deployment of machine learning in clinical prognostic modeling. As it is automated, AutoPrognosis can be used by clinical researchers to build prognostic models tailored to a given dataset without the need for in-depth knowledge of machine learning, clearing one of the most important hurdles to using these approaches in routine clinical practice.[39] Furthermore, this framework overcomes the "black-box" nature of machine learning models by converting the trained model into an interpretable and transparent risk equation.

AutoPrognosis automatically constructs an optimized prognostic model fit to the dataset at hand by tuning the parameters of an ensemble of state-of-the-art machine learning pipelines; each pipeline comprises an imputation algorithm, a feature processing algorithm, a machine learning prediction model, and a calibration algorithm. (Here, we deactivate the feature pre-processing module as the number of prognostic variables involved in model development is relatively small.) The overall Adjutorium model was constructed by fitting 10 binary classification ensemble models (optimized via AutoPrognosis) to predict outcomes at 10 distinct knots (time horizons spanning from 1 to 10 years from baseline, with 1-year increments). The AutoPrognosis algorithm creates this ensemble by tuning the parameters of the ML models using an advanced Bayesian optimization technique, and combining these tuned models using Bayesian model averaging.[19]

In order to convert the ML ensemble (created through Bayesian optimization) into a transparent model of risk, AutoPrognosis uses a symbolic regression methodology to *automatically* convert the trained ensemble model into an understandable mathematical equation that links patient variables to predicted outcomes. It does so using a search technique that optimizes parameterized symbolic expressions comprising combinations of uni-variate Meijer G-functions.[20] Survival curves were created by smoothing the coefficients for the symbolic expressions describing the model predictions at the 10 knots via cubic spline interpolation.

Cox model. A standard Cox proportional hazards (PH) model fit on the same data as Adjutorium was also assessed for comparison. Consistent with previous methods,[7] we applied two separate models, with different baseline hazards for ER positive and ER negative cancer. We included an age squared term to allow for non-linear effects of baseline age at diagnosis on breast cancer mortality. Tumor size and number of lymph nodes were both coded as continuous variables. Separate models where fit to each of the 10 imputed datasets,

and the resulting predictions of the 10 models (evaluated on validation data) where pooled using Rubin's rule.

The coefficients of the Cox PH model fitted to the training cohort (with breast cancer-specific outcomes) and averaged over the 10 imputed data sets are provided in Supplementary Table 1. The in-sample Harrell's concordance index of the pooled predictions for ER negative cancer was 0.72, whereas that for ER positive cancer was 0.80. HER2 status qualitatively interacts with ER status to modify risk of breast cancer mortality (HR for HER2 positive tumours is 0.73, 95% CI: 0.69-0.77 for patients with ER positive tumours, and 1.24, 95% CI: 1.20-1.28 for patients with ER negative tumours). This indicates that HER2 positive status is associated with reduced risk for mortality in ER negative cancer, but associates with relatively worse prognosis in ER positive cancer.

Model training. Patient samples from the NCRAS database were randomly split into two mutually exclusive cohorts: a training cohort of 316,690 patients used for model derivation, and an internal validation cohort of 79,172 patients used to evaluate model accuracy. The entire SEER cohort (571,635 patients) was reserved for external validation. We trained Adjutorium using the NCRAS data to predict breast cancer and all-cause mortality without adjuvant therapies by adjusting survival times for treatment effects, to create a counterfactual "untreated" survival cohort. Estimated survival time in absence of treatments was calculated as:

$$S_{bc}^{T=0} = S_{bc}^{T=1} \times HR, \tag{1}$$

where S_{bc} represents the uncensored survival time for each individual, T is the indicator for treatment, and HR is the hazard ratio associated with a specific treatment based on the EBCTCG meta-analysis.[21, 22] This is consistent with previous approaches used to create adjusted counterfactual survival times in cross-over trials.[40] The same procedure was applied to the Cox PH model. The Adjutorium model incorporates four treatments: chemotherapy, hormone therapy, bisphosphonates and trastuzumab. Other therapies, such as immunotherapy, targeted PARP and CDK4/6 inhibitors are primarily used for patients with metastatic cancer with no sufficient data on their usage as adjuvant therapies, hence we did not include them in our model.[41]

Model validation. We conducted internal and external validation of Adjutorium within the NCRAS validation cohort (n=79,172) and the SEER cohort (n=571,635), respectively. In addition, we also validate our model in the NCRAS-2 sub-cohort, which comprised 3,560 patients with complete treatment information. We validated predicted outcomes in the original unadjusted cohort, incorporating treatment effects for patients that had received therapy. Using this approach allowed us to evaluate the predictive accuracy of overall survival without treatment, and improvement of survival with treatment. As breast cancer mortality and mortality from other causes are competing causes, overall survival probability from all causes was calculated as follows:

$$P_{all}(t) = P_{bc}(t) \times P_{nbc}(t). \tag{2}$$

Here, $P_{all}(t)$, $P_{bc}(t)$ and $P_{nbc}(t)$ represent overall survival, survival from breast cancer, and survival from other non-breast cancer related causes at time horizon t, respectively. For individuals on adjuvant therapy, $P_{bc}(t)$ was calculated as a function of survival without treatment $P_{bc}^{T=0}(t)$ (as predicted by the trained model), and the effect of treatment, as follows:

$$P_{bc}^{T=1}(t) = \left(P_{bc}^{T=0}(t)\right)^{HR}. \tag{3}$$

Statistical analysis

Discriminative Accuracy. We compared the discriminative accuracy of Adjutorium in predicting all-cause and breast cancer-specific mortality at 3, 5 and 10 years from baseline relative to PREDICT v2.1[7] and the in-house Cox PH model fitted to the NCRAS training cohort. For the NCRAS-2 cohort, we only evaluated discriminative accuracy for 3- and 5-year outcomes since patients in this cohort were diagnosed in 2013, hence the maximum follow-up time in this cohort was less than 6 years. We assessed the discriminative accuracy of Adjutorium using the time-dependent area under receiver operating characteristic curve[25] (AUC-ROC), Harrell's concordance index[26] (C-index), and Uno's C-index.[27] Details on the mathematical definitions of each of these metrics can be found in Supplementary Information. For all evaluations, 95% confidence intervals were obtained using bootstrapped re-sampling of the validation data.

Calibration Accuracy. We evaluated the calibration curves of Adjutorium by comparing predicted risk of mortality with observed risk at the time horizons of interest. For each time horizon, we divided the risk ranges predicted by Adjutorium into 10 quantiles, and within each quantile, we estimated the observed risk in the corresponding patient samples using a Kaplan-Meier estimator.[43] Calibration curves were evaluated by plotting the predicted risks by Adjutorium on the x-axis, and the corresponding observed risk on the y-axis.

Sensitivity analyses. In order to examine the robustness of Adjutorium to missingness, we validated its performance separately on individuals with complete data and those with at least one missing variable. (In Supplementary Information, we also validate Adjutorium on individuals with different numbers of missing variables, and individuals with each variables missing.) Moreover, in order to assess the robustness of Adjutorium to time-cohort effects, due to changes in patient management and survival over time, we compared its discriminative accuracy with that of PREDICT in subsets of patients diagnosed within 1-year windows spanning from 2005 to 2016.

Subgroup analyses. We validated Adjutorium within specific patient subgroups stratified by age, ER status, HER2 status, tumour size and tumour grade. We specifically assessed the performance of Adjutorium relative to PREDICTv2.1 in patients aged more than 65 years, patients with larger tumours (>50 mm), and patients with negative ER status. Error counts (true positive and false positive cases, corresponding to the number of

cases misclassified) in each subgroup were obtained through decision thresholds that maximize the Youden J-statistic for each model.

Data and code availability

The data set used to derive and internally validate the model was obtained from the National Cancer Registration and Analysis Service. These data are held by Public Health England, and information on how to access these data can be found at http://ncin.org.uk/collecting_and_using_data/data_access. The data set used for external validation was obtained from the Surveillance, Epidemiology and Results program, which can be accessed at https://seer.cancer.gov/seertrack/data/request/. The code for the AutoPrognosis software is available at https://bitbucket.org/mvdschaar/mlforhealthlabpub.

Acknowledgments

The authors thank Prof. Eric Topol (Scripps research institute), Prof. David Dodwell (Oxford University), Prof. Marc Cullen (Stanford University) and Dr. Stephen Sammutt (Cambridge University) for their comments.

References

1. Fitzmaurice, C. *et al.* Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA oncology* **3**, 524–548 (2017).

2. Bray, F. *et al.* Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **68**, 394–424 (2018).

3. Guo, F., Kuo, Y.-f., Shih, Y. C. T., Giordano, S. H. & Berenson, A. B. Trends in breast cancer mortality by stage at diagnosis among young women in the u nited s tates. *Cancer* **124**, 3500–3509 (2018).

4. Sparano, J. A. *et al.* Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *The New England journal of medicine* (2019).

5. Symmans, W. F. *et al.* Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of Clinical Oncology* **25**, 4414–4422 (2007).

6. Wishart, G. C. *et al.* Predict: a new uk prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research* **12**, R1 (2010).

7. dos Reis, F. J. C. *et al.* An updated predict breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Research* **19**, 58 (2017).

8. Shachar, S. S. & Muss, H. B. Internet tools to enhance breast cancer care. *NPJ Breast Cancer* **2**, 16011 (2016).

9. Kattan, M. W. *et al.* American joint committee on cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA: a cancer journal for clinicians* **66**, 370–374 (2016).

10. National, G. A. U. Early and locally advanced breast cancer: diagnosis and management (2018).

11. van Maaren, M. C. *et al.* Validation of the online prediction tool predict v. 2.0 in the dutch breast cancer population. *European journal of cancer* **86**, 364–372 (2017).

12. Olivotto, I. A. *et al.* Population-based validation of the prognostic model adjuvant! for early breast cancer. *Journal of Clinical Oncology* **23**, 2716–2725 (2005).

13. Bhoo-Pathy, N. *et al.* Adjuvant! online is overoptimistic in predicting survival of asian breast cancer patients. *European Journal of Cancer* **48**, 982–989 (2012).

14. Campbell, H., Taylor, M., Harris, A. & Gray, A. An investigation into the performance of the adjuvant! online prognostic programme in early breast cancer for a cohort of patients in the united kingdom. *British journal of cancer* **101**, 1074 (2009).

15. Miao, H. *et al.* Validation of the cancermath prognostic tool for breast cancer in southeast asia. *BMC cancer* **16**, 820 (2016).

16. Ravdin, P. M. *et al.* Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of clinical oncology* **19**, 980–991 (2001).

17. Obermeyer, Z. & Emanuel, E. J. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* **375**, 1216 (2016).

18. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine* **376**, 2507 (2017).

19. Alaa, A. & Schaar, M. Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning. In *International Conference on Machine Learning*, 139–148 (2018).

20. Alaa, A. M. & van der Schaar, M. Demystifying black-box models with symbolic metamodels. In *Advances in Neural Information Processing Systems*, 11301–11311 (2019).

21. Group, E. B. C. T. C. *et al.* Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials. *The Lancet* **379**, 432–444 (2012).

22. Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. *New England journal of medicine* **353**, 1673–1684 (2005).

23. Alaa, A. M. & van der Schaar, M. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Scientific reports* **8**, 11242 (2018).

24. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. & van Der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants. *PloS one* **14**, e0213653 (2019).

25. Lambert, J. & Chevret, S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent roc curves. *Statistical methods in medical research* **25**, 2088–2102 (2016).

26. Harrell Jr, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15**, 361–387 (1996).

27. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* **30**, 1105–1117 (2011).

28. Noone, A. *et al.* Seer cancer statistics review, 1975-2015. *Bethesda, MD: National Cancer Institute* (2018).

29. Galea, M. H., Blamey, R. W., Elston, C. E. & Ellis, I. O. The nottingham prognostic index in primary breast cancer. *Breast cancer research and treatment* **22**, 207–219 (1992).

30. Michaelson, J. S. *et al.* Improved web-based calculators for predicting breast carcinoma outcomes. *Breast cancer research and treatment* **128**, 827–835 (2011).

31. Zhang, Z. Multiple imputation with multivariate imputation by chained equation (mice) package. *Annals of translational medicine* **4** (2016).

32. Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* **160**, 3–24 (2007).

33. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology* **5**, 412 (2014).

34. Down, S. K., Lucas, O., Benson, J. R. & Wishart, G. C. Effect of predict on chemotherapy/trastuzumab recommendations in her2-positive patients with early-stage breast cancer. *Oncology letters* **8**, 2757–2761 (2014).

35. Wishart, G. C. *et al.* Inclusion of ki67 significantly improves performance of the predict prognostication and prediction model for early breast cancer. *BMC cancer* **14**, 908 (2014).

36. Ács, B. *et al.* Ki-67 as a controversial predictive and prognostic marker in breast cancer patients treated with neoadjuvant chemotherapy. *Diagnostic pathology* **12**, 20 (2017).

37. Ware, J. H., Harrington, D., Hunter, D. J. & D'Agostino Sr, R. B. Missing data (2012).

38. Royston, P. Multiple imputation of missing values. *The Stata Journal* **4**, 227–241 (2004).

39. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* **25**, 44 (2019).

40. Latimer, N., Abrams, K. & Siebert, U. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. *BMC medical research methodology* **19**, 69 (2019).

41. Mayer, E. L. Targeting breast cancer with cdk inhibitors. *Current oncology reports* **17**, 20 (2015).

544 42. Lee, A. H. & Ellis, I. O. The nottingham prognostic index for invasive carcinoma of the breast.
545 *Pathology & Oncology Research* **14**, 113–115 (2008).

546 43. D'agostino, R. & Nam, B.-H. Evaluation of the performance of survival analysis models:
547 discrimination and calibration measures. *Handbook of statistics* **23**, 1–25 (2003).

Figures

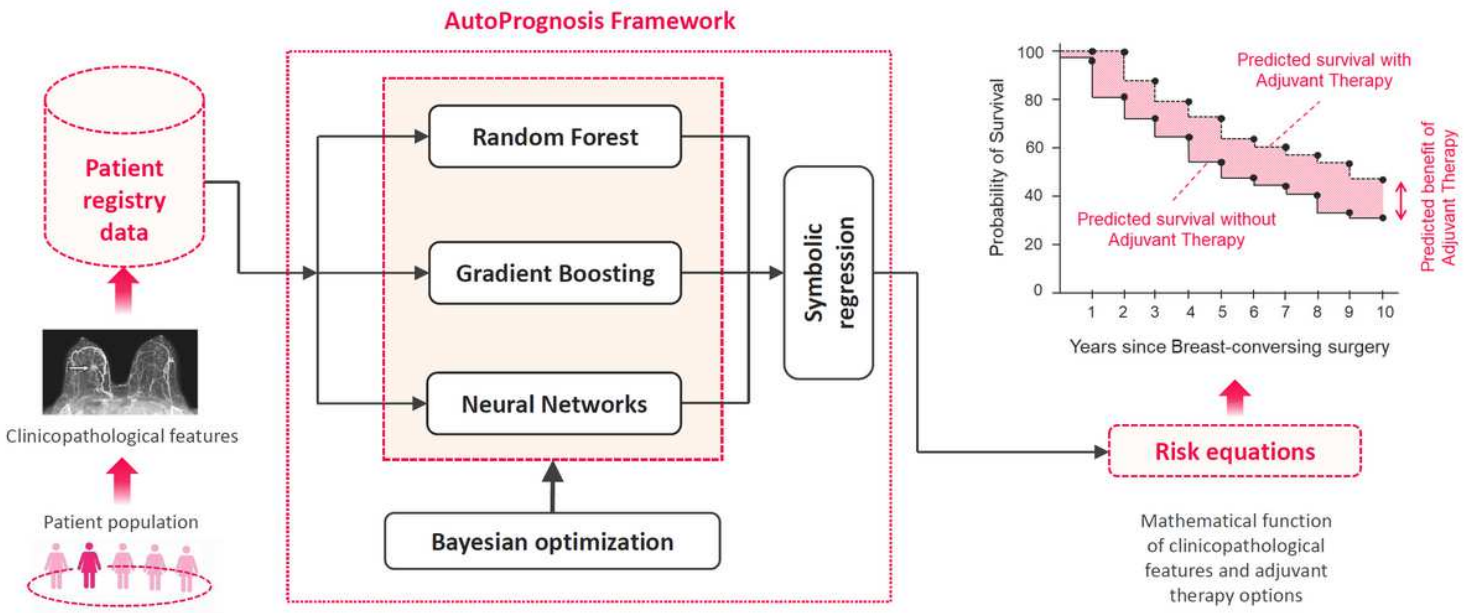


Figure 1

Schematic depiction of the AutoPrognosis framework. Given patient data, AutoPrognosis uses a Bayesian optimization algorithm to search for the optimal parameters of a collection of machine learning models and the optimal weight assigned to each model in an ensemble. (Here, we depict random forests, gradient boosting and neural network models as exemplary elements of the ensemble.) After fitting the ensemble model, a symbolic regression algorithm is used to convert the fitted model into a mathematical equation that maps patient variables to predicted risk. The end result is a mathematical equation that computes an individual patient's survival curve with and without a given therapy.

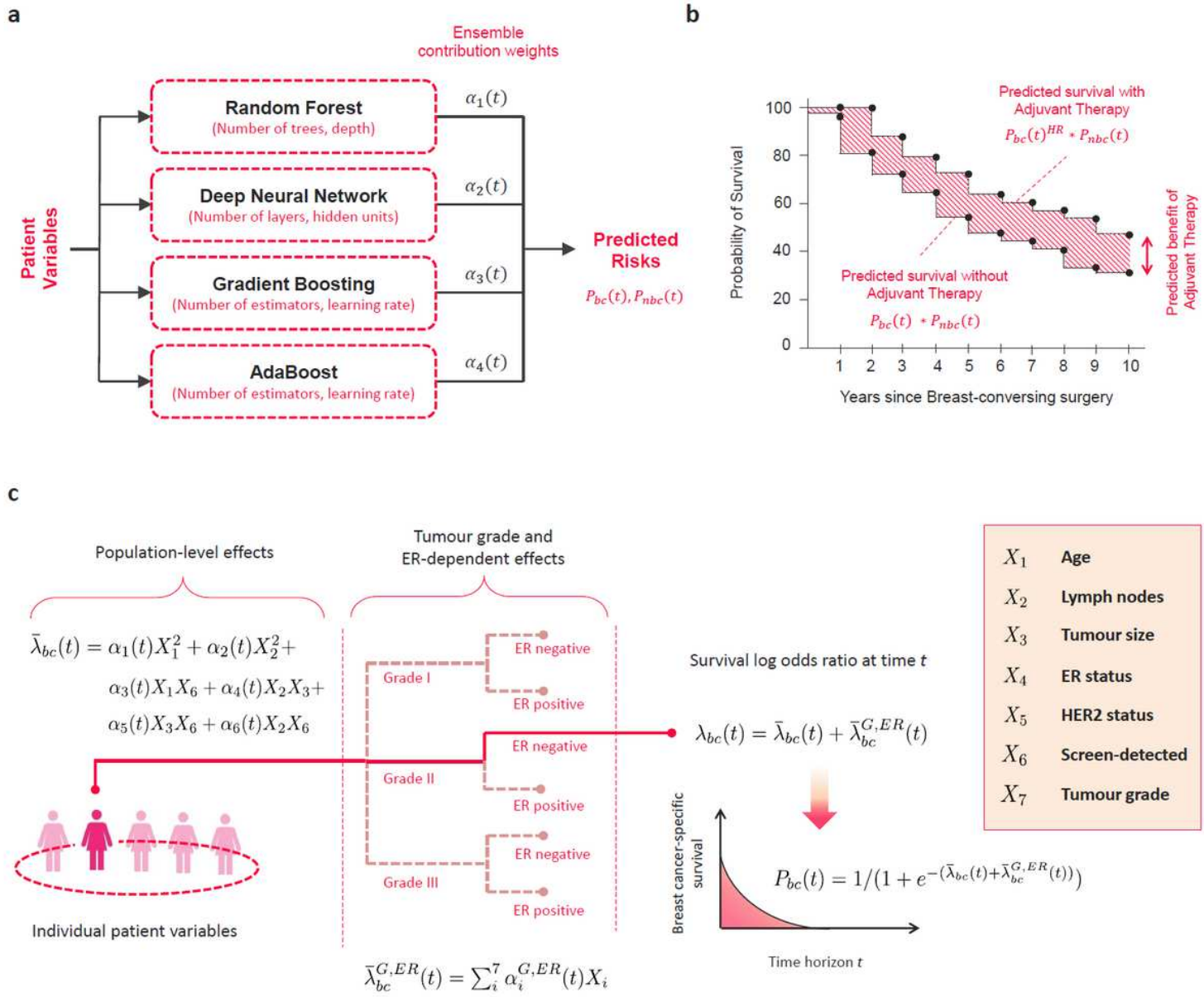


Figure 2

Illustration for the machine learning model underlying Adjutorium. a, The ensemble model learned by the AutoPrognosis software. The ensemble comprises four basic machine learning models: random forest, neural network, gradient boosting, and AdaBoost. The prediction issued by Adjutorium is a weighted combination of the predictions of the four members of the ensemble. Each model in the ensemble has a set of parameters (listed between brackets), and an assigned weight 𝛼(t) determining its contribution in the final prediction. Both the model parameters and its weight change depending on the prediction horizon t. Separate ensembles are trained to predict breast cancerspecific survival Pbc(t) and other cause survival Pnbc(t). b, The predicted survival curve for an exemplary patient (with and without adjuvant therapy). Here, each prediction horizon (1 to 10 years since diagnosis, with 1-year steps) corresponds to a knot in the survival curve, and each knot is associated with a distinct set of model parameters and contribution weights in the ensemble in a. c, Risk equations underlying Adjutorium as learned by the symbolic regression module in AutoPrognosis. Given the individual-level variables of a patient, the risk

equation evaluates the probability of survival at future time horizons. The log odds ratio for survival at time t comprises two components: (1) a population-level term that models non-linear effects of age and number of lymph nodes, in addition to interactions between different variables through six coefficients that are fixed for all patients, and (2) a tumour grade and ER-specific term that evaluates the linear effects of all prognostic factors with coefficients that are specific to every group of patients with the same grade and ER status. Here we show an exemplary patient with ER negative cancer and tumour grade 2 and. The risk equation is a mathematical abstraction for the predictions issued by the machine learning model in a.

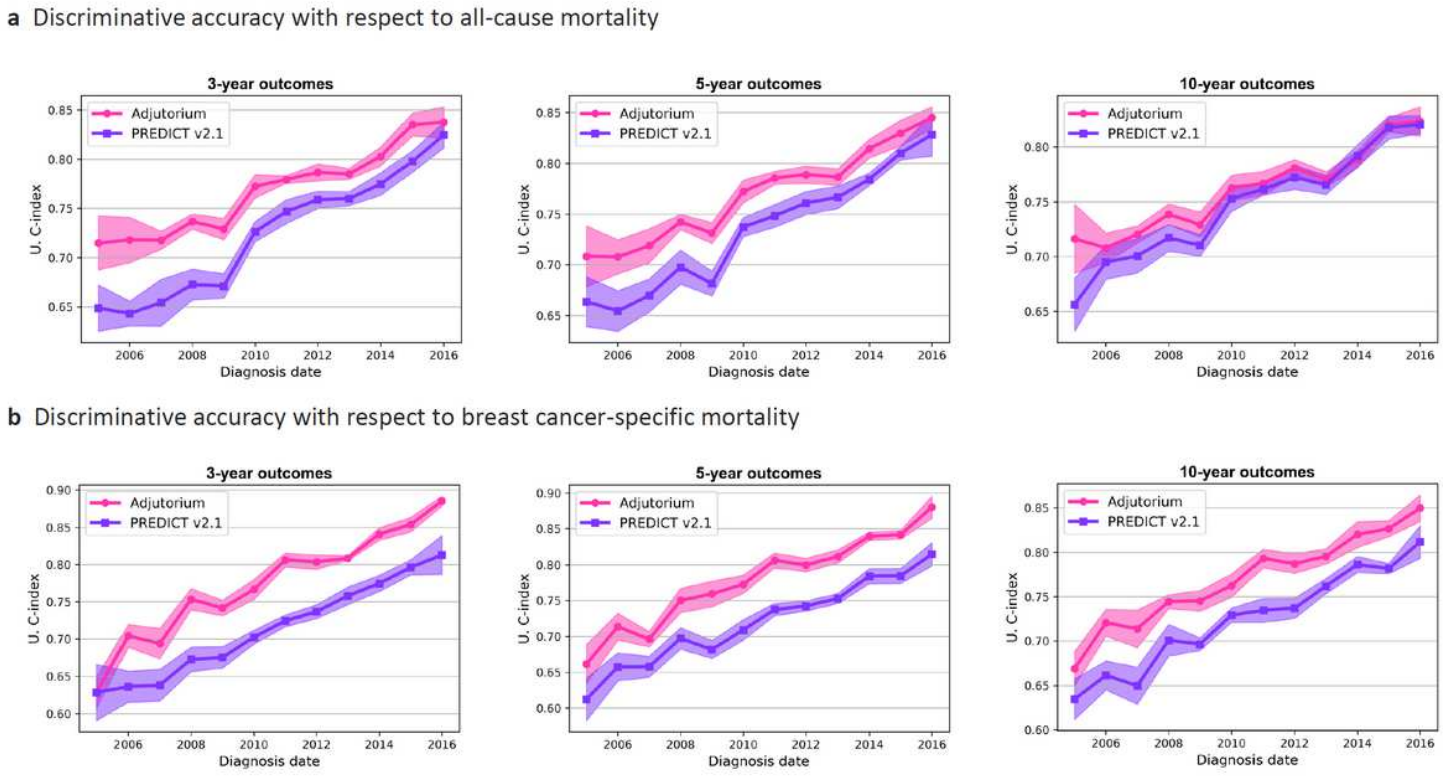
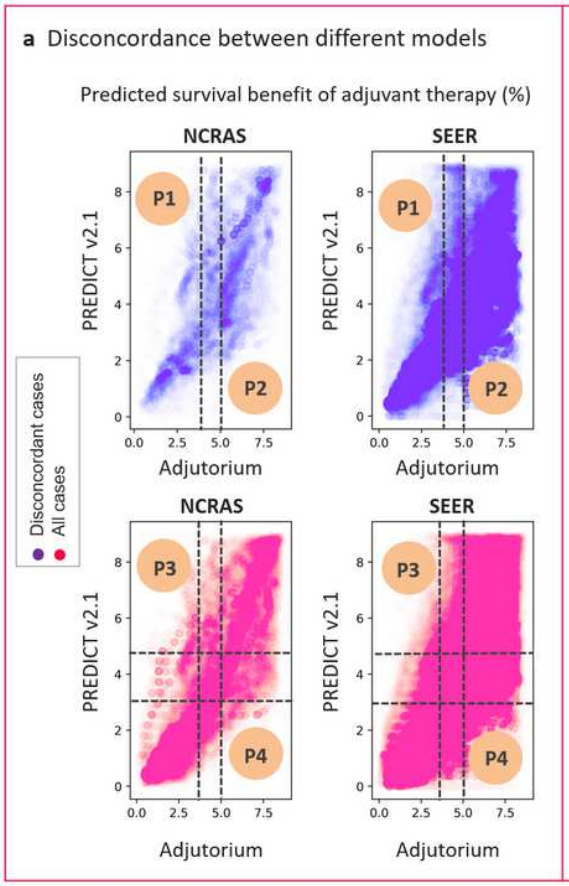


Figure 3

Discriminative accuracy evaluated in sub-cohorts of patients stratified by diagnosis date.



b Characteristics of patients in disconcordant cases

Population P1: untreated by Adjutorium, treated by MDT. **MR*:** NCRAS=0.85, SEER=0.88.

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=9,904)	55.7	2.5	22	10%	69%	41%	65%
SEER (n=73,866)	57.7	1.9	22	18%	80%	13%	___

Population P2: treated by Adjutorium, untreated by MDT. **MR:** NCRAS=1.13, SEER=1.08.

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=10,345)	63.0	5.6	28	2.7%	59%	37%	16.7%
SEER (n=101,861)	57.3	1.2	24	1.8%	41%	3%	___

Population P3: untreated by Adjutorium, treated by PREDICT. **MR:** NCRAS=0.92, SEER=0.75.

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=5,096)	60.0	7.9	23	6%	43%	47%	34%
SEER (n=12,105)	60.0	5.9	37	2%	29%	15%	___

Population P4: treated by Adjutorium, untreated by PREDICT. **MR:** NCRAS=1.20, SEER=1.60

	Age	Lymph nodes	Tumour size	Tumour grade I	ER positive	HER2 positive	Screen-detected
NCRAS (n=9,285)	58.2	1.9	25	1.9%	66%	20%	41%
SEER (n=99,389)	55.4	0.7	20	2.5%	54%	3.5%	___

* MR stands for "mortality ratio", defined as the ratio between the 10-year mortality rate in the selected population and that of the overall population.

Figure 4

Comparison between therapeutic decisions informed by Adjutorium and PREDICT v2.1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NMIsupplementaryinformation.pdf
- NMITable1.pdf
- NMITable2.pdf