Machine learning techniques enhance conventional methods in predicting clinical outcomes

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-004007
Article Type:	Research
Date Submitted by the Author:	13-Sep-2013
Complete List of Authors:	Gupta, Sunil; Deakin University, Centre for Pattern Recognition and Data Analytics Tran, Truyen; Deakin University, Centre for Pattern Recognition and Data Analytics Luo, Wei; Deakin University, Centre for Pattern Recognition and Data Analytics Phung, Dinh; Deakin University, Centre for Pattern Recognition and Data Analytics Kennedy, Richard; Deakin University, School of Medicine Broad, Adam; Barwon Health, Andrew Love Cancer Centre Campbell, David; Barwon Health, Andrew Love Cancer Centre Kipp, David; Barwon Health, Andrew Love Cancer Centre Singh, Madhu; Barwon Health, Andrew Love Cancer Centre Khasraw, Mustafa; Barwon Health, Andrew Love Cancer Centre Matheson, Leigh; Barwon Southwest Integrated Cancer Service, Ashley, David; Barwon Health, Andrew Love Cancer Centre; Deakin University, School of Medicine Venkatesh, Svetha; Deakin University, Centre for Pattern Recognition and Data Analytics
<b>Primary Subject Heading</b>:	Oncology
Secondary Subject Heading:	Health informatics
Keywords:	Cancer, Survival, Prediction, Machine Learning, Electronic Medical Record

Machine learning techniques enhance conventional methods in predicting clinical outcomes

Sunil Gupta[1], Truyen Tran[1, 2], Wei Luo[1], Dinh Phung[1], Richard Lee Kennedy[3], Adam Broad[4], David Campbell [4], David Kipp [4], Madhu Singh[4], Mustafa Khasraw [3,4], Leigh Matheson[5],David M Ashley[3,4,5], Svetha Venkatesh[1]*

[1]Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia

[2]Department of Computing, Curtin University, Perth, Western Australia, Australia

[3]School of Medicine, Deakin University, Geelong, Victoria, Australia

[4]Andrew Love Cancer Centre, Barwon Health, Geelong, Victoria, Australia

[5]Barwon Southwest Integrated Cancer Service, Geelong, Victoria, Australia

*Correspondence to: Svetha Venkatesh, Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia 3220 svetha.venkatesh@deakin.edu.au

Telephone: +61 3 5227 2905 Fax: +61 3 5227 2028

Total Number of Text Pages: 24

Total Number of Tables: 6

Total Number of Illustrations: 0

Acknowledgements: The authors have no financial disclosures, conflicts of interest, or external funding sources to claim.

Abstract word count:	293
Word count:	4,680

Keywords: Cancer; Survival; Prediction; Machine learning; Electronic medical record

Abstract

Objectives: Using the prediction of cancer outcome as a model, we have tested the hypothesis that through analysing routinely collected digital data contained in an electronic administrative record (EAR), using machine learning techniques, we could enhance conventional methods in predicting clinical outcomes.

Setting: A regional cancer centre in Australia.

Participants: Disease specific data from a purpose built cancer registry (ECO) from 869 patients was used to predict survival at 6, 12, and 24 months. The model was validated with data from a further 94 patients, and results compared to assessment of five specialist oncologists. Machine-learning prediction using ECO data was compared with that using EAR and a model combining ECO and EAR data.

Primary and secondary outcome measures: Survival prediction accuracy in terms of the area of the ROC curve.

Results: The ECO model yielded AUCs of 0·87 (95% CI=0·848–0·890) at six months, 0·796 (95% CI=0·774–0·823) at 12 months, and 0·764 (95% CI=0·737–0·789) at 24 months. Each was slightly better than the performance of the clinician panel. The model performed consistently across a range of cancers, including rare cancers. Combining ECO and EAR data yielded better prediction than the ECO-based model (AUCs ranging from 0·757 to 0·997 for 6 months, AUCs from 0·689 to 0·988 for 12 months, and AUCS from 0·713 to 0·973 for 24

months). The best prediction was for genitourinary, head and neck, lung, skin and upper gastrointestinal tumours.

Conclusion: Machine learning applied to information from a disease specific (cancer) database and the EAR can be used to predict clinical outcomes. Importantly, the approach described made use of a digital data that is already routinely collected but under-exploited by clinical health systems.

Strengths and limitations of this study

- This is the first study using machine learning of both administrative and registry data for cancer survival prediction.
- A single prognosis model is produced across all cancers, improving prediction accuracy on rare cancers.
- This is a retrospective study in a single centre.

Introduction

Over the past two decades there has been an explosion in the use of digital footprints to monitor and predict human behaviours. The source of data used for this purpose is our on-line use of the internet, the emails we send and transactions we make. Analysis of these footprints through machine learning techniques (MLT) have been exploited in the public domain by government and business to predict behaviours and inform investment decisions. In research MLT have also been used to analyse gene expression data, [15, 23] and for medical image analysis, [24, 25]. However, to date, there has been little exploration of these methodologies in the clinical setting. We hypothesised that MLT may offer a paradigm shift in clinical medicine that can address core issues with large and complex datasets. These techniques offer the potential to derive adaptive systems from diverse datasets, discover latent connections between data items, and to predict outcomes.

Most hospitals routinely collect large digital electronic administrative records (EAR). These are primarily used for organisational financial management. Historically, they have not been used extensively for clinical or research purposes. If these large data sets are able to be exploited using MLT it may open the way to optimise the use of collected administrative data to assist in predicting patients outcome, planning individualised patient care, monitoring resource utilisation, and improving institutional performance. [8, 9] The accurate assessment of comorbid status would improve assessment of prognosis and guide treatment decisions. [10-13] Other important information that may be contained or inferred

from an EAR includes geographical and demographic data, socioeconomic status, and history of health care facility utilisation. [14-16]

In this study, using cancer outcome prediction as a model, we wished to test the hypothesis that routinely collected digital health data, if analysed by state of the art, validated, machine learning techniques could be used to assist conventional tools in predicting clinical outcomes.

Accurate prediction of survival in patients with cancer remains a challenge due to the ever-increasing heterogeneity and complexity of cancer, treatment options, and patient populations. If achieved, reliable predictions could assist personalised care and treatment, and improve institutional performance in cancer management. In current practice clinicians use data collected at the bedside in consultations, medical records or purpose built cancer registries to aid prognostication and decision making.

The notion of using MLT to predict cancer prognosis from clinical and pathological data is not a new one. [1, 2] However, with the advent of more sophisticated and better validated techniques, not only is more accurate prediction possible, but the range of data incorporated into decision aids can be increased. [3-5]. The need to improve cancer care systems by creating linkages between registries and epidemiological surveillance through analysis of complex and large clinical databases has recently been highlighted. [6, 7]

In this study we tested the capability of MLT to predict patient outcomes in a heterogeneous cohort of cancer patients. We have interrogated two data sets: the first a

purpose-built cancer specific registry (ECO) containing demographic and tumour-related data items according to an Australian nationally agreed protocol; the second a hospital digital data set containing information about the patient’s previous admissions and presentations (EAR). Finally, in a test group of 94 patients, we examined the performance of machine-learning methods in aiding a panel of expert clinicians in predicting patient survival.

Patients and Methods

Study design

This is a retrospective study using the electronic administrative record (EAR) and a specialised cancer registry (ECO) from Barwon Health, the only public tertiary institution in a region of Australia with more than 350,000 residents. With a unified hospital identity number in use across the region, Barwon Health's EAR provides a single point of access for information on patient encounters with the health system, including hospitalizations, ED visits, medications, and treatments. In addition, the Andrew Love Cancer Centre at Barwon health has a specialised cancer registry called ECO, which captures clinical data for patients in the region. ECO records information on demographics, primary tumour and metastatic tumour, cancer stage, tumour size, lymph nodes, and breast tumour specific information. Treatment type, outcomes, including death, and recurrence information (primary and metastatic) are also recorded. Table 1 shows the variables used for survival prediction. The cohort for this study consists of 963 patients identified in ECO who were first diagnosed in year 2009. Among these patients, 736 patients also had records in the EAR. Ethics approval was obtained from the Hospital and Research Ethics Committee at Barwon Health (number 12/83). Deakin University has reciprocal ethics authorization with Barwon Health.

Analyses

The analyses centred on predicting cancer survival since the date of diagnosis, defined as the date of tumour resection. Each patient was a unit of observation in the predictive problem: Patient data collected prior to the diagnosis date were used to construct the independent variables; Survival status in a period following the assessment was the dependent variable. Two analyses were performed: The first compared survival prediction made by machine learning models and the clinician panel, based on only information from ECO. The second analysis evaluated the added discriminative power provided by EAR, by comparing the best machine learning models using three sets of predicting variables: variables from ECO (Table 1), variables from EAR (appendix), and the union of the two.

Comparing predictions by machine learning models and clinician

In the first analysis, all 963 patients in the ECO registry were randomly divided into a derivation cohort of 869 patients and a validation cohort of 94 patients (Table 2). To collect clinician prediction, patients in the validation cohort were assigned to a panel of five oncologists for survival prediction. For each patient, the oncologist was asked to estimate the survival probabilities based on the independent variables in Table 1. All clinicians estimated the patient's survival status by producing a probability for each of the three time periods—6 months, 1 year, and 2 years. When making this assessment the clinicians did not have knowledge of the treatment type offered or given to the patient. Three machine-learning models were trained on the derivation cohort using the same set of independent variables, one for each prediction period. Each of the machine learning models was an ensemble of 400 support vector machines [17] with linear kernel (i.e., the output of the model was the average of 400 support-vector-machine outputs). Each of the support vector machines was trained using a random 80% subsampling (without replacement) of the derivation cohort.[18] Two measures were taken to improve the training process. First, to compensate for the imbalance between the two outcomes (there were more survivals than deaths), we oversampled the non-surviving cases by 50% in each training subsample. Next, variable selection was performed through fitting a generalized linear model with elastic net regularization [19] (alpha parameter set to 0·1 and lambda parameter selected using 5-fold internal cross-validation) and variables with zero coefficients were removed. After the machine learning models were constructed, they were applied to predict survival probabilities for each patient in the validation cohort. Both the clinician and model predictions were validated with the actual outcomes in the ECO registry. Prediction performance was measured using the area under the ROC curve (AUC), also known as the C-statistic. [20] 95% confidence intervals of AUCs were computed using 1000 bootstrap samples of validation cohort.

Comparing discriminative information from specialized registry and routine data

The second analysis compared the discriminative power of two data sources (ECO and EAR). In this analysis, clinician predictions were not solicited. Among the 869 patients in the derivation subset of Cohort 1, only 664 have records in the EAR and these patients were included in the second analysis (Cohort 2, Table 2). Survival prediction models were derived based on three sets of independent variables: 1) independent variables from EAR (EAR only); 2) independent variables from ECO (ECO only); 3) the union of the two sets (*EAR + ECO*). Similar to the previous analysis, the models were trained using random 80% subsamples and the modelling process was identical. However, the models were evaluated not using the validation cohort. Instead, for each 80% subsample, the remaining 20% was used to compute the AUC and its 95% confidence interval.

The Wilcoxon rank-sum test was applied to answer the following comparison problems:

1. Does *ECO only* provide more discriminative power than *EAR only*?
2. Does *EAR + ECO* provide more discriminative power than *EAR only*?
3. Does *EAR + ECO* provide more discriminative power than *ECO only*?

Details of the machine learning model and the predictor variables can be found in the Appendix.

Results

The cohorts for the two analyses are summarized in Table 2. The comparison between the algorithmic predictions and the clinician predictions are summarized in Table 3. The model had comparable performance to that of the clinicians, with the performance of the machine learning model marginally better (AUC ranging from 0·76 to 0·87) than that of the clinicians (AUC ranging from 0·75 to 0·79) for all three prediction periods. This similarity in accuracy between algorithmic predictions and the clinician predictions was observed across different cancer types. Consider the predictions for six-month survival. Out of 15 breast cancer cases, the clinicians made 15 correct predictions and the algorithm made 14; Out of 18 lung cancer cases, the clinicians made 13 correct predictions and the algorithm made 14; Out of 7 haematological cases, both the clinicians and the algorithm made all predictions correctly.

Similar results were observed on 12-month and 24-month survival predictions for different cancers.

Prediction of 6-month survival using the three models is shown in Table 4. There were no deaths from breast cancer during this period. Comparing the ECO model with the EAR model, AUCs were comparable for colorectal, genitourinary, haematological, head and neck and skin tumours. The EAR model was significantly better ($p < 0{\cdot}05$) for rare tumours; CNS, upper gastrointestinal and unassigned primary source tumours. For each tumour type, the model using both ECO and EAR data yielded similar or better performance to the models using information from only one of the two databases. AUCs for the combined model ranged from 0·76 to 1·0. The combined data model showed particularly improved performance over ECO data (p value <·05) for all tumour streams except and breast and CNS tumours.

Data for 12-month survival prediction is shown in Table 5. Cancer-specific ECO data yielded better prediction than EAR data ($p < 0{\cdot}05$) for gynaecological, haematological lung, skin and unknown primary cancers. Otherwise, ECO and EAR models yielded generally similar results. The model using combined data performed better than EAR (p value < ·05) for all tumour streams other than CNS, head and neck and upper gastro tumours. The model using combined data was better than ($P < 0{\cdot}05$) ECO for all cancers except breast, CNS, gynaecological and haematological cancers.

Table 6 shows data for 24-month survival prediction by the three models. The ECO model yielded superior prediction (p value < ·05) to the EAR model for breast, genitourinary, gynaecological, lung, skin and unknown primary cancers, while the EAR model was superior to the ECO model for haematological and head and neck tumours. Once more the model that performed the best was that derived from both ECO and EAR data with AUCs ranging from 0·71 to 0·97 across the range of cancers and particularly enhanced performance for all cancers except breast, colorectal, gynaecological and unknown primary tumours compared to the ECO. In summary, over all time periods, the performance of the combined model was better than ECO (p < 0·05) for genitourinary, head and neck, lung, skin and upper gastrointestinal tumours.

Discussion

In this study, using cancer outcome prediction as a model, we wished to test the hypothesis that routinely collected digital health data, if analysed by MLT, could be used to assist conventional tools in predicting clinical outcomes.

Applying machine learning to data from the electronic administrative record (EAR) alone predicted clinical outcomes with reasonable accuracy. Using the purpose built ECO data set, the predictive tool also performed well across a broad range of cancer types, and in both cases the predictive accuracies were at least as good as that of a panel of five expert clinicians. Importantly a predictive tool derived from both the purpose built clinical registry and administrative data had even greater predictive ability.

The wealth of administrative data contained in the EAR includes information on comorbid conditions and previous clinic and hospital attendances as well as a drug history. There is considerable potential to use this data to improve clinical care across a spectrum of diseases. [8, 9]

We have designed this study as retrospective and in a single centre; it will be of major interest to observe how it performs in a variety of settings. The number of cases used to assess performance of the models is relatively small. The strengths include the comparison of machine learning tools with expert clinical opinion and the fact that very detailed and well-validated data was available both directly related to the cancer and that contained in the EAR. The generic nature of this approach makes it unnecessary to generate separate predictive models for different types of cancer. This was a particular advantage for rarer forms of cancer where predications using more conventional methods are very challenging.

Predictive tools derived from clinical data items have considerable potential to improve clinical care, but must be suitably optimised and shown to perform equally well in diverse clinical settings. [21, 22] Clinical databases have become more widely available and increasingly complex in recent years. The extent and complexity of data available to clinicians means that novel approaches to managing data and supporting clinical decisions are needed. Machine learning approaches can not only cope with complex datasets, but also adapt in real time and across different clinical settings.

The approach used in this study offers superior performance to previous machine learning approaches to predicting cancer survival. [1-5] Previous models have been derived for single cancer types, or for a limited range of cancers. The model described here performed well across a wide range of cancers. One advantage of this generic approach may be the ability to predict outcomes in less common cancers where limited data might preclude development of specific models. The fact that our model derived from administrative and cancer-related data performed slightly better than a panel of expert clinicians not only validates the potential utility of the model but suggests that it may be useful in assessing quality of care and also in settings where specialist care is not available.

Clinical outcomes in any illness are determined not only by specific factors related to the illness itself but also by the patient’s general state of health and by the presence of other chronic medical conditions often coded in an EAR if the individual traffics the health service.[10-13] As well, a particularly novel and important aspect of the use historical data from the EAR in machine learning is that it effectively captures the health care institutions current and previous performance. These data can be applied to any individual entering the system with a newly diagnosed cancer, as we have modelled here. As well they could also be used for quality and performance monitoring.

In conclusion, machine learning applied to information from a disease specific (cancer) database and the EAR can be used to predict outcomes. Improved prediction of outcome has the potential to help clinicians make more meaningful decisions about treatment and to assist with planning of future social and care needs. Most importantly, the approach described makes use of digital data that is already routinely collected but under-exploited by clinical health systems.

References

1. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, Jr., Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*. 1997;79:857-862
2. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. *Oncology*. 1999;57:281-286
3. Manilich EA, Kiran RP, Radivoyevitch T, Lavery I, Fazio VW, Remzi FH. A novel data-driven prognostic model for staging of colorectal cancer. *Journal of the American College of Surgeons*. 2011;213:579-588, 588.e571-572
4. Gao P, Zhou X, Wang ZN, Song YX, Tong LL, Xu YY, Yue ZY, Xu HM. Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. The tnm staging system. *PLoS ONE [Electronic Resource]*. 2012;7:e42015
5. Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW. Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer*. 2012;15:230-238
6. Johnson CJ, Weir HK, Fink AK, German RR, Finch JL, Rycroft RK, Yin D, Accuracy of Cancer Mortality Study G. The impact of national death index linkages on population-based cancer survival rates in the united states. *Cancer Epidemiology*. 2013;37:20-28
7. Khoury MJ, Lam TK, Ioannidis JP, Hartge P, Spitz MR, Buring JE, Chanock SJ, Croyle RT, Goddard KA, Ginsburg GS, Herceg Z, Hiatt RA, Hoover RN, Hunter DJ, Kramer BS, Lauer MS, Meyerhardt JA, Olopade OI, Palmer JR, Sellers TA, Seminara D, Ransohoff DF, Rebbeck TR, Tourassi G, Winn DM, Zauber A, Schully SD. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology, Biomarkers & Prevention*. 2013;22:508-516
8. Appari A, Eric Johnson M, Anthony DL. Meaningful use of electronic health record systems and process quality of care: Evidence from a panel data analysis of u.S. Acute-care hospitals. *Health Services Research*. 2013;48:354-375
9. Fitzhenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, Reeves RM, Aronsky D, Elkin PL, Messina VP, Speroff T. Exploring the frontier of electronic health record surveillance: The case of postoperative complications. *Medical Care*. 2013;51:509-516
10. Lund L, Borre M, Jacobsen J, Sorensen HT, Norgaard M. Impact of comorbidity on survival of danish prostate cancer patients, 1995-2006: A population-based cohort study. *Urology*. 2008;72:1258-1262
11. Tetsche MS, Norgaard M, Jacobsen J, Wogelius P, Sorensen HT. Comorbidity and ovarian cancer survival in denmark, 1995-2005: A population-based cohort study. *International Journal of Gynecological Cancer*. 2008;18:421-427
12. Lieffers JR, Baracos VE, Winget M, Fassbender K. A comparison of charlson and elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data. *Cancer*. 2011;117:1957-1965
13. Braithwaite D, Moore DH, Satariano WA, Kwan ML, Hiatt RA, Kroenke C, Caan BJ. Prognostic impact of comorbidity among long-term breast cancer survivors: Results from the lace study. *Cancer Epidemiology, Biomarkers & Prevention*. 2012;21:1115-1125
14. Jones LE, Doebbeling CC. Beyond the traditional prognostic indicators: The impact of primary care utilization on cancer survival. *Journal of Clinical Oncology*. 2007;25:5793-5799

15. Chang CM, Su YC, Lai NS, Huang KY, Chien SH, Chang YH, Lian WC, Hsu TW, Lee CC. The combined effect of individual and neighborhood socioeconomic status on cancer survival rates. *PLoS ONE [Electronic Resource]*. 2012;7:e44325

16. Sant M, Minicozzi P, Allemani C, Cirilli C, Federico M, Capocaccia R, Budroni M, Candela P, Crocetti E, Falcini F, Ferretti S, Fusco M, Giacomin A, La Rosa F, Mangone L, Natali M, Leon MP, Traina A, Tumino R, Zambon P. Regional inequalities in cancer care persist in italy and can influence survival. *Cancer Epidemiology*. 2012;36:541-547

17. Cortes C, Vapnik V. Support vector machine. *Machine learning*. 1995;20:273-297

18. Politis D, Romano J, Wolf M. *Subsampling*. Springer-Verlag, New York; 1999.

19. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 2010;33:1

20. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*. 2005;27:83-85

21. Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. *BMC Medical Research Methodology*. 2012;12:102

22. Chen HC, Chen JJ. Assessment of reproducibility of cancer survival risk predictions across medical centers. *BMC Medical Research Methodology*. 2013;13:25

23. Zhao X, Rodland EA, Sorlie T, Naume B, Langerod A, Frigessi A, Kristensen VN, Borresen-Dale AL, Lingjaerde OC. Combining gene signatures improves prediction of breast cancer survival. *PLoS ONE [Electronic Resource]*. 2011;6:e17845

24. Li C, Zhang S, Zhang H, Pang L, Lam K, Hui C, Zhang S. Using the k-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. *Computational & Mathematical Methods in Medicine*. 2012;2012:876545

25. Huang ML, Hung YH, Lee WM, Li RK, Wang TH. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *Journal of Medical Systems*. 2012;36:407-414

26. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, Mani S, Levy MA, Dai Q, Denny JC. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA ... Annual Symposium Proceedings/AMIA Symposium*. 2011;2011:1564-1572

Table 1: ECO variables used for survival prediction

patient demographics
post code
Gender
Age
tumour characteristics
primary site (in ICD-10 code)
tumour stream
morphology (in ICD-O-3 code)
histologic grade
metastatic sites
most valid basis of diagnosis
performance status diagnosis
stage basis (pathological or clinical)
stage (TNM)
tumour size
nodes taken
positive nodes
breast cancer related variables
oestrogen receptor
progesterone receptor
human epidermal growth factor receptor 2 (HER2)

Table 2: Characteristics of Derivation and Validation Cohorts

	Cohort 1: ECO		**Cohort 2: ECO and EAR (n=664)**
	Derivation (n=869)	**Validation (n=**94**)**	
Age (SD)	67·6 (14·6)	68·4 (13·6)	66·3(14·9)
Gender: Males	487 *	48	381
Tumour stream			
Genitourinary	172	21	135
Colorectal	140	14	115
Lung	121	18	96
Breast	122	15	74
Haematological	99	7	85
Upper gastro	83	9	57
Skin	36	1	28
Head and Neck	35	0	30
Gynaecological	19	4	17
CNS	15	1	9

Unknown primary	38	9	26

[*] 2 unspecified

Table 3: Performance of survival prediction: comparison between machine learning method and clinicians

Survival Period	**AUC (95% CI)**	
	Clinician panel	**Machine learning model**
6 months	0·79 (0·76, 0·81)	0·87 (0·85, 0·89)
1 year	0·79 (0·76, 0·81)	0·80 (0·77, 0·82)
2 years	0·75 (0·73, 0·78)	0·76 (0·74, 0·79)

Table 4: Prediction performance of machine learning algorithms: 6-month survival

Cancer type	**Area Under ROC Curve (95% CI)**		
	EAR only	**ECO only**	**EAR + ECO**
Genitourinary	·81 (·77, ·85)	·82 (·78, ·86)	·88 (·85, ·91) [*,†]
Colorectal	·84 (·80, ·88)	·85 (·81, ·89)	·88 (·84, ·91) [*,†]
Lung	·71 (·67, ·76)	·73 (·69, ·77) [*]	·77 (·73, ·82) [*,†]
Breast		no deaths in the period	
Haematological	·73 (·68, ·79)	·74 (·69, ·79)	·76 (·71, ·81)
Upper gastro	·74 (·69, ·78)	·64 (·60, ·69)	·84 (·80, ·87) [†]
Skin	·84 (·77, ·90)	·85 (·79, ·91)	·91 (·86, ·96) [*,†]
Head and neck	·66 (·61, ·71)	·70 (·64, ·75)	·77 (·72, ·82) [*,†]
Gynaecological	·97 (·94, ·99)	·99 (·98, 1·0) [*]	1·0 (·99, 1·0) [*]
CNS	·89 (·85, ·94)	·84 (·78, 0·90)	·82 (·77, ·88)
Unknown primary	·92 (·89, 95)	·79 (·75, ·84)	·90 (·87, ·93) [*,†]

[*] Significantly greater than *EAR only*. [†] Significantly greater than *ECO only*.

Table 5: Prediction performance of machine learning algorithms: 12-month survival

	Area Under ROC Curve (95% CI)		
Cancer type	***EAR only***	***ECO only***	***EAR + ECO***
Genitourinary	·79 (·75, ·83)	·79 (·75, ·83)	·84 (·80, ·87) *,†
Colorectal	·82 (·78, ·86)	·83 (·79, ·86)	·87 (·83, ·90) *,†
Lung	·73 (·69, ·77)	·78 (·73, ·82) *	·82 (·78, ·86) *,†
Breast	·71 (·65, ·78)	·90 (·86, ·94)	·92 (·89, ·96) *
Haematological	·63 (·59, ·68)	·70 (·66, ·75) *	·69 (·64, ·74) *
Upper gastro	·62 (·57, ·66)	·70 (·65, ·74) *	·72 (·68, ·76) *
Skin	·76 (·71, ·88)	·89 (·85, ·93) *	·93 (·90, ·96) *
Head and neck	·77 (·73, ·88)	·68 (·63, 73)	·79 (·75, ·84) †
Gynaecological	·95 (·92, ·97)	1·0 (1·0, 1·0) *	·99 (·98, 1·0) *
CNS	·66 (·58, ·73)	·68 (·61, ·76)	·69 (·63, ·76)
Unknown primary	·87 (·84, ·91)	·81 (·77, ·85)	·88 (·84, ·91)

*Significantly greater than *EAR only*. †Significantly greater than *ECO only*.

Table 6: Prediction performance of machine learning algorithms: 24-month survival

	Area Under ROC Curve (AUC)		
Cancer type	**EAR only**	**ECO only**	**EAR + ECO**
Genitourinary	·73 (·69, ·78)	·84 (·81, ·88) *	·86 (·82, ·89) *,†
Colorectal	·76 (·72, 80)	·76 (·72, ·80)	·76 (·72, ·80)
Lung	·74 (·69, ·78)	·78 (·73, ·82) *	·82 (·79, ·86) *,†
Breast	·67 (·61, ·73)	·86 (·82, ·90) *	·88 (·84, ·92) *
Haematological	·73 (·68, ·77)	·70 (·66, ·75)	·80 (·76, ·84) *,†
Upper gastro	·81 (·77, ·85)	·77 (·72, ·81)	·87 (·83, ·90) *,†
Skin	·71 (·65, ·76)	·85 (·80, ·89) *	·94 (·92, ·97) *,†
Head and neck	·74 (·70, ·78)	·66 (·51, ·61)	·71 (·67, ·76) †
Gynaecological	·96 (·94, ·99)	·99 (·98, 1·0) *	·97 (·95, ·99)
CNS	·83 (·78, ·89)	·87 (·82, ·93)	·96 (·93, ·99) *,†
Unknown primary	·74 (·70, ·79)	·78 (·74, ·82) *	·80 (·76, ·84) *

*Significantly greater than *EAR only*. †Significantly greater than *ECO only*.

Appendix

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

Derivation of the machine learning model

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

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

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

Predictors for the machine learning models

Table 1 EMR-based predictors

demographics

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

cancer specific diagnoses

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

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

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

Table 2 ECO-based predictors.

patient demographics

- Gender
- Age

tumour characteristics

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

breast cancer related variables

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

References

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

STARD checklist for reporting of studies of diagnostic accuracy

(version January 2003)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	1-4
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	5-8
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	9
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	9
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	Yes
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	retrospective
Test methods	7	The reference standard and its rationale.	9-10
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	9-10
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	9-10
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	N/A
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	N/A
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	10-12
	13	Methods for calculating test reproducibility, if done.	10-12
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	9
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	21
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	N/A
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	N/A
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	N/A
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	N/A
	20	Any adverse events from performing the index tests or the reference standard.	N/A
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	23-25
	22	How indeterminate results, missing data and outliers of the index tests were handled.	N/A
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	N/A
	24	Estimates of test reproducibility, if done.	N/A
DISCUSSION	25	Discuss the clinical applicability of the study findings.	15-17

Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-004007.R1
Article Type:	Research
Date Submitted by the Author:	17-Feb-2014
Complete List of Authors:	Gupta, Sunil; Deakin University, Centre for Pattern Recognition and Data Analytics Tran, Truyen; Deakin University, Centre for Pattern Recognition and Data Analytics Luo, Wei; Deakin University, Centre for Pattern Recognition and Data Analytics Phung, Dinh; Deakin University, Centre for Pattern Recognition and Data Analytics Kennedy, Richard; Deakin University, School of Medicine Broad, Adam; Barwon Health, Andrew Love Cancer Centre Campbell, David; Barwon Health, Andrew Love Cancer Centre Kipp, David; Barwon Health, Andrew Love Cancer Centre Singh, Madhu; Barwon Health, Andrew Love Cancer Centre Khasraw, Mustafa; Barwon Health, Andrew Love Cancer Centre Matheson, Leigh; Barwon Southwest Integrated Cancer Service, Ashley, David; Barwon Health, Andrew Love Cancer Centre; Deakin University, School of Medicine Venkatesh, Svetha; Deakin University, Centre for Pattern Recognition and Data Analytics
<b>Primary Subject Heading</b>:	Oncology
Secondary Subject Heading:	Health informatics
Keywords:	Cancer, Survival, Prediction, Machine Learning, Electronic Medical Record

Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry

Sunil Gupta[1], Truyen Tran[1, 2], Wei Luo[1], Dinh Phung[1], Richard Lee Kennedy[3], Adam Broad[4], David Campbell [4], David Kipp [4], Madhu Singh[4], Mustafa Khasraw [3,4], Leigh Matheson[5],David M Ashley[3,4,5], Svetha Venkatesh[1]*

[1]Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia

[2]Department of Computing, Curtin University, Perth, Western Australia, Australia

[3]School of Medicine, Deakin University, Geelong, Victoria, Australia

[4]Andrew Love Cancer Centre, Barwon Health, Geelong, Victoria, Australia

[5]Barwon Southwest Integrated Cancer Service, Geelong, Victoria, Australia

*Correspondence to: Svetha Venkatesh, Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia 3220 svetha.venkatesh@deakin.edu.au

Telephone: +61 3 5227 2905 Fax: +61 3 5227 2028

Total Number of Text Pages: 24

Total Number of Tables: 6

Total Number of Illustrations: 0

Abstract word count: 293

Word count: 4,680

Keywords: Cancer; Survival; Prediction; Machine learning; Electronic medical record

Abstract

Objectives: Using the prediction of cancer outcome as a model, we have tested the hypothesis that through analysing routinely collected digital data contained in an electronic administrative record (EAR), using machine learning techniques, we could enhance conventional methods in predicting clinical outcomes.

Setting: A regional cancer centre in Australia.

Participants: Disease specific data from a purpose built cancer registry (ECO) from 869 patients was used to predict survival at 6, 12, and 24 months. The model was validated with data from a further 94 patients, and results compared to assessment of five specialist oncologists. Machine-learning prediction using ECO data was compared with that using EAR and a model combining ECO and EAR data.

Primary and secondary outcome measures: Survival prediction accuracy in terms of the area of the ROC curve.

Results: The ECO model yielded AUCs of 0·87 (95% CI=0·848–0·890) at six months, 0·796 (95% CI=0·774–0·823) at 12 months, and 0·764 (95% CI=0·737–0·789) at 24 months. Each was slightly better than the performance of the clinician panel. The model performed consistently across a range of cancers, including rare cancers. Combining ECO and EAR data yielded better prediction than the ECO-based model (AUCs ranging from 0·757 to 0·997 for 6 months, AUCs from 0·689 to 0·988 for 12 months, and AUCS from 0·713 to 0·973 for 24

months). The best prediction was for genitourinary, head and neck, lung, skin and upper gastrointestinal tumours.

Conclusion: Machine learning applied to information from a disease specific (cancer) database and the EAR can be used to predict clinical outcomes. Importantly, the approach described made use of a digital data that is already routinely collected but under-exploited by clinical health systems.

Strengths and limitations of this study

- This is the first study using machine learning of both administrative and registry data for cancer survival prediction.
- A single prognosis model is produced across all cancers, improving prediction accuracy on rare cancers.
- This is a retrospective study in a single centre.

Introduction

Over the past two decades there has been an explosion in the use of digital footprints to monitor and predict human behaviours. The source of data used for this purpose is our on-line use of the internet, the emails we send and transactions we make. Analysis of these footprints through machine learning techniques (MLT) have been exploited in the public domain by government and business to predict behaviours and inform investment decisions. In research MLT have also been used to analyse gene expression data, [1, 2] and for medical image analysis. [3, 4] However to date, there has been little exploration of these methodologies in the clinical setting. We hypothesised that MLT may offer a paradigm shift in clinical medicine that can address core issues with large and complex datasets. These techniques offer the potential to derive adaptive systems from diverse datasets, discover latent connections between data items, and to predict outcomes.

Most hospitals routinely collect large digital electronic administrative records (EAR). These are primarily used for organisational financial management. Historically, they have not been used extensively for clinical or research purposes. If these large data sets are able to be exploited using MLT it may open the way to optimise the use of collected administrative data to assist in predicting patients outcome, planning individualised patient care, monitoring resource utilisation, and improving institutional performance. [5, 6] The accurate assessment of comorbid status would improve assessment of prognosis and guide treatment decisions. [7-10] Other important information that may be contained or inferred

from an EAR includes geographical and demographic data, socioeconomic status, and history of health care facility utilisation. [2, 11, 12]

In this study, using cancer outcome prediction as a model, we wished to test the hypothesis that routinely collected digital health data, if analysed by state of the art, validated, machine learning techniques could be used to assist conventional tools in predicting clinical outcomes.

Accurate prediction of survival in patients with cancer remains a challenge due to the ever-increasing heterogeneity and complexity of cancer, treatment options, and patient populations. If achieved, reliable predictions could assist personalised care and treatment, and improve institutional performance in cancer management. In current practice clinicians use data collected at the bedside in consultations, medical records or purpose built cancer registries to aid prognostication and decision making.

The notion of using MLT to predict cancer prognosis from clinical and pathological data is not a new one. [13, 14] However, with the advent of more sophisticated and better validated techniques, not only is more accurate prediction possible, but the range of data incorporated into decision aids can be increased. [15-17]. The need to improve cancer care systems by creating linkages between registries and epidemiological surveillance through analysis of complex and large clinical databases has recently been highlighted. [18, 19]

In this study we tested the capability of MLT to predict patient outcomes in a heterogeneous cohort of cancer patients. We have interrogated two data sets: the first a

purpose-built cancer specific registry (ECO) containing demographic and tumour-related data items according to an Australian nationally agreed protocol; the second a hospital digital data set containing information about the patient's previous admissions and presentations (EAR). Finally, in a test group of 94 patients, we examined the performance of machine-learning methods in aiding a panel of expert clinicians in predicting patient survival.

Patients and Methods

Study design

This is a retrospective study using the electronic administrative record (EAR) and a specialised cancer registry (ECO) from Barwon Health, the only public tertiary institution in a region of Australia with more than 350,000 residents. With a unified hospital identity number in use across the region, Barwon Health's EAR provides a single point of access for information on patient encounters with the health system, including hospitalizations, ED visits, medications, and treatments. In addition, the Andrew Love Cancer Centre at Barwon health has a specialised cancer registry called ECO, which captures clinical data for patients in the region. ECO records information on demographics, primary tumour and metastatic tumour, cancer stage, tumour size, lymph nodes, and breast tumour specific information. Treatment type, outcomes, including death, and recurrence information (primary and metastatic) are also recorded. Table 1 shows the variables used for survival prediction. The cohort for this study consists of 963 patients identified in ECO who were first diagnosed in year 2009. The study completion date was October 31, 2012; therefore all patients had at least 2 year and 10 months follow-up. Among these patients, 736 patients also had records in the EAR. Ethics approval was obtained from the Hospital and Research Ethics Committee at Barwon Health (number 12/83). Deakin University has reciprocal ethics authorization with Barwon Health.

Analyses

The analyses centred on predicting cancer survival since the date of diagnosis, defined as the date of tumour resection. Each patient was a unit of observation in the predictive problem: Patient data collected prior to the diagnosis date were used to construct the independent variables; Survival status in a period following the assessment was the dependent variable. Two analyses were performed: The first compared survival prediction made by machine learning models and the clinician panel, based on only information from ECO. The second analysis evaluated the added discriminative power provided by EAR, by comparing the best machine learning models using three sets of predicting variables: variables from ECO (Table 1), variables from EAR (appendix), and the union of the two.

Although a survival analysis model (e.g., a proportional hazards model [20]) is commonly used in modelling risk factors, such models are not designed to predict events. In this study, survival was directly modelled using classification models to optimize prediction accuracy.

Comparing predictions by machine learning models and clinician

In the first analysis, all 963 patients in the ECO registry were randomly divided into a derivation cohort of 869 patients and a validation cohort of 94 patients (Table 2). To collect clinician prediction, patients in the validation cohort were assigned to a panel of five oncologists for survival prediction. For each patient, the oncologist was asked to estimate the survival probabilities based on the independent variables in Table 1. All clinicians estimated the patient's survival status by producing a probability for each of the three time periods—6 months, 1 year, and 2 years. When making this assessment the clinicians did not have knowledge of the treatment type offered or given to the patient. Three machine-learning models were trained on the derivation cohort using the same set of independent variables, one for each prediction period. Each of the machine learning models was an ensemble of 400 support vector machines [21] with linear kernel (i.e., the output of the model was the average of 400 support-vector-machine outputs in Platt's a posteriori probabilities[22]). Ensemble was used to control the variability introduced by $L1$ feature selection. Each of the support vector machines was trained using a random 80% subsampling (without replacement) of the derivation cohort.[23] The soft margin parameter (C) of SVM was selected through cross-validation. Two measures were taken to improve the training process. First, to compensate for the imbalance between the two outcomes (there were more survivals than deaths), we oversampled the non-surviving cases by 50% in each training subsample. Next, variable selection was performed through fitting a generalized linear model with elastic net regularization [24] (alpha parameter set to 0·1 and lambda parameter selected using 5-fold internal cross-validation) and variables with zero coefficients were removed. After the machine learning models were constructed, they were applied to predict survival probabilities for each patient in the validation cohort. Both the clinician and model predictions were validated with the actual outcomes in the ECO registry. Prediction performance was measured using the area under the ROC curve (AUC), also known as the C-statistic. [25] 95% confidence intervals of AUCs were computed using 1000 bootstrap samples of validation cohort.

Comparing discriminative information from specialized registry and routine data

The second analysis compared the discriminative power of two data sources (ECO and EAR). In this analysis, clinician predictions were not solicited. Among the 869 patients in the derivation subset of Cohort 1, only 664 have records in the EAR and these patients were included in the second analysis (Cohort 2, Table 2). Survival prediction models were derived based on three sets of independent variables: 1) independent variables from EAR (EAR only); 2) independent variables from ECO (ECO only); 3) the union of the two sets (*EAR + ECO*). Similar to the previous analysis, the models were trained using 400 random subsamples comprising 80% data of the cohort-2 and the modelling process was identical. However, the models were evaluated not using the validation cohort. Instead, for each 80% subsample, the remaining 20% was used to compute the AUC and its 95% confidence interval.

The Wilcoxon rank-sum test was applied to answer the following comparison problems:

1. Does *ECO only* provide more discriminative power than *EAR only*?
2. Does *EAR + ECO* provide more discriminative power than *EAR only*?
3. Does *EAR + ECO* provide more discriminative power than *ECO only*?

Details of the machine learning model and the predictor variables can be found in the Appendix.

Results

The cohorts for the two analyses are summarized in Table 2. The comparison between the algorithmic predictions and the clinician predictions are summarized in Table 3. The model had comparable performance to that of the clinicians, with the performance of the machine learning model marginally better (AUC ranging from 0·76 to 0·87) than that of the clinicians (AUC ranging from 0·75 to 0·79) for all three prediction periods. This similarity in accuracy between algorithmic predictions and the clinician predictions was observed across different cancer types. Consider the predictions for six-month survival. Out of 15 breast cancer cases, the clinicians made 15 correct predictions and the algorithm made 14; Out of 18 lung cancer cases, the clinicians made 13 correct predictions and the algorithm made 14; Out of 7 haematological cases, both the clinicians and the algorithm made all predictions correctly. Similar results were observed on 12-month and 24-month survival predictions for different cancers.

Prediction of 6-month survival using the three models is shown in Table 4. There were no deaths from breast cancer during this period. Comparing the ECO model with the EAR model, AUCs were comparable for colorectal, genitourinary, haematological, head and neck and skin tumours. The EAR model was significantly better ($p < 0·05$) for rare tumours; CNS, upper gastrointestinal and unassigned primary source tumours. For each tumour type, the model using both ECO and EAR data yielded similar or better performance to the models using information from only one of the two databases. AUCs for the combined model ranged from 0·76 to 1·0. The combined data model showed particularly improved performance over ECO data (p value <·05) for all tumour streams except and breast and CNS tumours.

Data for 12-month survival prediction is shown in Table 5. Cancer-specific ECO data yielded better prediction than EAR data (p < 0·05) for gynaecological, haematological lung, skin and unknown primary cancers. Otherwise, ECO and EAR models yielded generally similar results. The model using combined data performed better than EAR (p value < ·05) for all tumour streams other than CNS, head and neck and upper gastro tumours. The model using combined data was better than (P < 0·05) ECO for all cancers except breast, CNS, gynaecological and haematological cancers.

Table 6 shows data for 24-month survival prediction by the three models. The ECO model yielded superior prediction (p value < ·05) to the EAR model for breast, genitourinary, gynaecological, lung, skin and unknown primary cancers, while the EAR model was superior to the ECO model for haematological and head and neck tumours. Once more the model that performed the best was that derived from both ECO and EAR data with AUCs ranging from 0·71 to 0·97 across the range of cancers and particularly enhanced performance for all cancers except breast, colorectal, gynaecological and unknown primary tumours compared to the ECO. In summary, over all time periods, the performance of the combined model was better than ECO (p < 0·05) for genitourinary, head and neck, lung, skin and upper gastrointestinal tumours.

One of the key advantage of using machine learning technique is that it can combine the large number of non-clinical factors with the few clinical risk factors. In this study, the model selected most of the known clinical risk factors including *patient age*, *cancer staging*, *performance status,* and *tumour size*. In addition it also found some useful non-clinical risk factors, including the type of the last hospital admission (emergency vs. elective), the frequency of ED visits within the previous 3 months and 6 months (related to both cancer and other medical conditions).

Discussion

In this study, using cancer outcome prediction as a model, we wished to test the hypothesis that routinely collected digital health data, if analysed by MLT, could be used to assist conventional tools in predicting clinical outcomes.

Applying machine learning to data from the electronic administrative record (EAR) alone predicted clinical outcomes with reasonable accuracy. Using the purpose built ECO data set, the predictive tool also performed well across a broad range of cancer types, and in both cases the predictive accuracies were at least as good as that of a panel of five expert clinicians. Importantly a predictive tool derived from both the purpose built clinical registry and administrative data had even greater predictive ability.

The wealth of administrative data contained in the EAR includes information on comorbid conditions and previous clinic and hospital attendances as well as a drug history. There is considerable potential to use this data to improve clinical care across a spectrum of diseases. [5, 6]

Most patients in the study were followed up for 3 years, which may not be adequate to capture all oncologic outcomes, especially for those cancers with low mortality rate. We have designed this study as retrospective and in a single centre; it will be of major interest to observe how it performs in a variety of settings. The number of cases used to assess performance of the models is relatively small. The strengths include the comparison of machine learning tools with expert clinical opinion and the fact that very detailed and well-validated data was available both directly related to the cancer and that contained in the EAR. The generic nature of this approach makes it unnecessary to generate separate predictive models for different types of cancer. This was a particular advantage for rarer forms of cancer where predications using more conventional methods are very challenging.

Predictive tools derived from clinical data items have considerable potential to improve clinical care, but must be suitably optimised and shown to perform equally well in diverse clinical settings. [26, 27] Clinical databases have become more widely available and increasingly complex in recent years. The extent and complexity of data available to clinicians means that novel approaches to managing data and supporting clinical decisions are needed. Machine learning approaches can not only cope with complex datasets, but also adapt in real time and across different clinical settings.

The approach used in this study offers superior performance to previous machine learning approaches to predicting cancer survival. [13-17] Previous models have been derived for single cancer types, or for a limited range of cancers. The model described here performed well across a wide range of cancers. One advantage of this generic approach may be the ability to predict outcomes in less common cancers where limited data might preclude development of specific models. The fact that our model derived from administrative and cancer-related data performed slightly better than a panel of expert clinicians not only validates the potential utility of the model but suggests that it may be useful in assessing quality of care and also in settings where specialist care is not available. An alternative approach to borrow information across different cancer types is call multi-task learning. We are currently exploring this approach as well.

Clinical outcomes in any illness are determined not only by specific factors related to the illness itself but also by the patient’s general state of health and by the presence of other chronic medical conditions often coded in an EAR if the individual traffics the health service.[7-10] As well, a particularly novel and important aspect of the use historical data from the EAR in machine learning is that it effectively captures the health care institutions current and previous performance. These data can be applied to any individual entering the system with a newly diagnosed cancer, as we have modelled here. As well they could also be used for quality and performance monitoring.

In conclusion, machine learning applied to information from a disease specific (cancer) database and the EAR can be used to predict outcomes. Improved prediction of outcome has the potential to help clinicians make more meaningful decisions about treatment and to assist with planning of future social and care needs. Most importantly, the approach described makes use of digital data that is already routinely collected but under-exploited by clinical health systems.

Acknowledgements: The authors have no financial disclosures, conflicts of interest, or external funding sources to claim.

Contributorship Statement: All authors of this research paper have directly participated in the planning, execution, or analysis of the study. All authors of this paper have read and approved the final version submitted.

Competing Interests: None

Data Sharing Statement: No additional data

References

1. Zhao X, Rodland EA, Sorlie T, et al. Combining gene signatures improves prediction of breast cancer survival. *PLoS ONE [Electronic Resource]*. 2011;6:e17845
2. Chang CM, Su YC, Lai NS, et al. The combined effect of individual and neighborhood socioeconomic status on cancer survival rates. *PLoS ONE [Electronic Resource]*. 2012;7:e44325
3. Li C, Zhang S, Zhang H, et al. Using the k-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. *Computational & Mathematical Methods in Medicine*. 2012;2012:876545
4. Huang ML, Hung YH, Lee WM, et al. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *Journal of Medical Systems*. 2012;36:407-414
5. Appari A, Eric Johnson M, Anthony DL. Meaningful use of electronic health record systems and process quality of care: Evidence from a panel data analysis of u.S. Acute-care hospitals. *Health Services Research*. 2013;48:354-375
6. Fitzhenry F, Murff HJ, Matheny ME, et al. Exploring the frontier of electronic health record surveillance: The case of postoperative complications. *Medical Care*. 2013;51:509-516
7. Lund L, Borre M, Jacobsen J, et al. Impact of comorbidity on survival of danish prostate cancer patients, 1995-2006: A population-based cohort study. *Urology*. 2008;72:1258-1262
8. Tetsche MS, Norgaard M, Jacobsen J, et al. Comorbidity and ovarian cancer survival in denmark, 1995-2005: A population-based cohort study. *International Journal of Gynecological Cancer*. 2008;18:421-427
9. Lieffers JR, Baracos VE, Winget M, et al. A comparison of charlson and elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data. *Cancer*. 2011;117:1957-1965
10. Braithwaite D, Moore DH, Satariano WA, et al. Prognostic impact of comorbidity among long-term breast cancer survivors: Results from the lace study. *Cancer Epidemiology, Biomarkers & Prevention*. 2012;21:1115-1125
11. Jones LE, Doebbeling CC. Beyond the traditional prognostic indicators: The impact of primary care utilization on cancer survival. *Journal of Clinical Oncology*. 2007;25:5793-5799
12. Sant M, Minicozzi P, Allemani C, et al. Regional inequalities in cancer care persist in italy and can influence survival. *Cancer Epidemiology*. 2012;36:541-547
13. Burke HB, Goodman PH, Rosen DB, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*. 1997;79:857-862
14. Lundin M, Lundin J, Burke HB, et al. Artificial neural networks applied to survival prediction in breast cancer. *Oncology*. 1999;57:281-286
15. Manilich EA, Kiran RP, Radivoyevitch T, et al. A novel data-driven prognostic model for staging of colorectal cancer. *Journal of the American College of Surgeons*. 2011;213:579-588, 588.e571-572

16. Gao P, Zhou X, Wang ZN, et al. Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. The tnm staging system. *PLoS ONE [Electronic Resource]*. 2012;7:e42015

17. Kim W, Kim KS, Lee JE, et al. Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer*. 2012;15:230-238

18. Johnson CJ, Weir HK, Fink AK, et al .Accuracy of Cancer Mortality Study G. The impact of national death index linkages on population-based cancer survival rates in the united states. *Cancer Epidemiology*. 2013;37:20-28

19. Khoury MJ, Lam TK, Ioannidis JP, et al. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology, Biomarkers & Prevention*. 2013;22:508-516

20. Cox DR, Oakes D. *Analysis of survival data*. CRC Press; 1984.

21. Cortes C, Vapnik V. Support vector machine. *Machine learning*. 1995;20:273-297

22. Lin H-T, Lin C-J, Weng RC. A note on platt's probabilistic outputs for support vector machines. *Mach. Learn.* 2007;68:267-276

23. Politis D, Romano J, Wolf M. *Subsampling*. Springer-Verlag, New York; 1999.

24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 2010;33:1

25. Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*. 2005;27:83-85

26. Chen HC, Kodell RL, Cheng KF, et al. Assessment of performance of survival prediction models for cancer prognosis. *BMC Medical Research Methodology*. 2012;12:102

27. Chen HC, Chen JJ. Assessment of reproducibility of cancer survival risk predictions across medical centers. *BMC Medical Research Methodology*. 2013;13:25

Table 1: ECO variables used for survival prediction

patient demographics
post code
Gender
Age
tumour characteristics
primary site (in ICD-10 code)
tumour stream
morphology (in ICD-O-3 code)
histologic grade
metastatic sites
most valid basis of diagnosis
performance status diagnosis
stage basis (pathological or clinical)
stage (TNM)
tumour size
nodes taken
positive nodes
breast cancer related variables
oestrogen receptor
progesterone receptor
human epidermal growth factor receptor 2 (HER2)

Table 2: Characteristics of Derivation and Validation Cohorts

	Cohort 1: ECO		**Cohort 2: ECO and EAR (n=664)**
	Derivation (n=869)	**Validation (n=94)**	
Age (SD)	67·6 (14·6)	68·4 (13·6)	66·3(14·9)
Gender: Males	487 *	48	381
Tumour stream			
Genitourinary	172	21	135
Colorectal	140	14	115
Lung	121	18	96
Breast	122	15	74
Haematological	99	7	85
Upper gastro	83	9	57
Skin	36	1	28
Head and Neck	35	0	30
Gynaecological	19	4	17
CNS	15	1	9

Unknown primary	38	9	26

[*] 2 unspecified

Table 3: Performance of survival prediction: comparison between machine learning method and clinicians

Survival Period	**AUC (95% CI)**	
	Clinician panel	**Machine learning model**
6 months	0·79 (0·76, 0·81)	0·87 (0·85, 0·89)
1 year	0·79 (0·76, 0·81)	0·80 (0·77, 0·82)
2 years	0·75 (0·73, 0·78)	0·76 (0·74, 0·79)

Table 4: Prediction performance of machine learning algorithms: 6-month survival

	Area Under ROC Curve (95% CI)		
Cancer type	**EAR only**	**ECO only**	**EAR + ECO**
Genitourinary	·81 (·77, ·85)	·82 (·78, ·86)	·88 (·85, ·91) [*,†]
Colorectal	·84 (·80, ·88)	·85 (·81, ·89)	·88 (·84, ·91) [*,†]
Lung	·71 (·67, ·76)	·73 (·69, ·77) [*]	·77 (·73, ·82) [*,†]
Breast		no deaths in the period	
Haematological	·73 (·68, ·79)	·74 (·69, ·79)	·76 (·71, ·81)
Upper gastro	·74 (·69, ·78)	·64 (·60, ·69)	·84 (·80, ·87) [†]
Skin	·84 (·77, ·90)	·85 (·79, ·91)	·91 (·86, ·96) [*,†]
Head and neck	·66 (·61, ·71)	·70 (·64, ·75)	·77 (·72, ·82) [*,†]
Gynaecological	·97 (·94, ·99)	·99 (·98, 1·0) [*]	1·0 (·99, 1·0) [*]
CNS	·89 (·85, ·94)	·84 (·78, 0·90)	·82 (·77, ·88)
Unknown primary	·92 (·89, 95)	·79 (·75, ·84)	·90 (·87, ·93) [*,†]

[*]Significantly greater than *EAR only*. [†]Significantly greater than *ECO only*.

Table 5: Prediction performance of machine learning algorithms: 12-month survival

	Area Under ROC Curve (95% CI)		
Cancer type	***EAR only***	***ECO only***	***EAR + ECO***
Genitourinary	·79 (·75, ·83)	·79 (·75, ·83)	·84 (·80, ·87) [*,†]
Colorectal	·82 (·78, ·86)	·83 (·79, ·86)	·87 (·83, ·90) [*,†]
Lung	·73 (·69, ·77)	·78 (·73, ·82) [*]	·82 (·78, ·86) [*,†]
Breast	·71 (·65, ·78)	·90 (·86, ·94)	·92 (·89, ·96) [*]
Haematological	·63 (·59, ·68)	·70 (·66, ·75) [*]	·69 (·64, ·74) [*]
Upper gastro	·62 (·57, ·66)	·70 (·65, ·74) [*]	·72 (·68, ·76) [*]
Skin	·76 (·71, ·88)	·89 (·85, ·93) [*]	·93 (·90, ·96) [*]
Head and neck	·77 (·73, ·88)	·68 (·63, 73)	·79 (·75, ·84) [†]
Gynaecological	·95 (·92, ·97)	1·0 (1·0, 1·0) [*]	·99 (·98, 1·0) [*]
CNS	·66 (·58, ·73)	·68 (·61, ·76)	·69 (·63, ·76)
Unknown primary	·87 (·84, ·91)	·81 (·77, ·85)	·88 (·84, ·91)

[*] Significantly greater than *EAR only*. [†] Significantly greater than *ECO only*.

Table 6: Prediction performance of machine learning algorithms: 24-month survival

	Area Under ROC Curve (AUC)		
Cancer type	**EAR only**	**ECO only**	**EAR + ECO**
Genitourinary	·73 (·69, ·78)	·84 (·81, ·88) *	·86 (·82, ·89) *,†
Colorectal	·76 (·72, 80)	·76 (·72, ·80)	·76 (·72, ·80)
Lung	·74 (·69, ·78)	·78 (·73, ·82) *	·82 (·79, ·86) *,†
Breast	·67 (·61, ·73)	·86 (·82, ·90) *	·88 (·84, ·92) *
Haematological	·73 (·68, ·77)	·70 (·66, ·75)	·80 (·76, ·84) *,†
Upper gastro	·81 (·77, ·85)	·77 (·72, ·81)	·87 (·83, ·90) *,†
Skin	·71 (·65, ·76)	·85 (·80, ·89) *	·94 (·92, ·97) *,†
Head and neck	·74 (·70, ·78)	·66 (·51, ·61)	·71 (·67, ·76) †
Gynaecological	·96 (·94, ·99)	·99 (·98, 1·0) *	·97 (·95, ·99)
CNS	·83 (·78, ·89)	·87 (·82, ·93)	·96 (·93, ·99) *,†
Unknown primary	·74 (·70, ·79)	·78 (·74, ·82) *	·80 (·76, ·84) *

*Significantly greater than *EAR only*. †Significantly greater than *ECO only*.

Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry

Sunil Gupta[1], Truyen Tran[1, 2], Wei Luo[1], Dinh Phung[1], Richard Lee Kennedy[3], Adam Broad[4], David Campbell [4], David Kipp [4], Madhu Singh[4], Mustafa Khasraw [3,4], Leigh Matheson[5],David M Ashley[3,4,5], Svetha Venkatesh[1]*

[1]Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia

[2]Department of Computing, Curtin University, Perth, Western Australia, Australia

[3]School of Medicine, Deakin University, Geelong, Victoria, Australia

[4]Andrew Love Cancer Centre, Barwon Health, Geelong, Victoria, Australia

[5]Barwon Southwest Integrated Cancer Service, Geelong, Victoria, Australia

*Correspondence to: Svetha Venkatesh, Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia 3220 svetha.venkatesh@deakin.edu.au

Telephone: +61 3 5227 2905 Fax: +61 3 5227 2028

Total Number of Text Pages: 24

Total Number of Tables: 6

Total Number of Illustrations: 0

Acknowledgements: The authors have no financial disclosures, conflicts of interest, or external funding sources to claim.

Abstract word count: 293

Word count: 4,680

Keywords: Cancer; Survival; Prediction; Machine learning; Electronic medical record

Abstract

Objectives: Using the prediction of cancer outcome as a model, we have tested the hypothesis that through analysing routinely collected digital data contained in an electronic administrative record (EAR), using machine learning techniques, we could enhance conventional methods in predicting clinical outcomes.

Setting: A regional cancer centre in Australia.

Participants: Disease specific data from a purpose built cancer registry (ECO) from 869 patients was used to predict survival at 6, 12, and 24 months. The model was validated with data from a further 94 patients, and results compared to assessment of five specialist oncologists. Machine-learning prediction using ECO data was compared with that using EAR and a model combining ECO and EAR data.

Primary and secondary outcome measures: Survival prediction accuracy in terms of the area of the ROC curve.

Results: The ECO model yielded AUCs of 0·87 (95% CI=0·848–0·890) at six months, 0·796 (95% CI=0·774–0·823) at 12 months, and 0·764 (95% CI=0·737–0·789) at 24 months. Each was slightly better than the performance of the clinician panel. The model performed consistently across a range of cancers, including rare cancers. Combining ECO and EAR data yielded better prediction than the ECO-based model (AUCs ranging from 0·757 to 0·997 for 6 months, AUCs from 0·689 to 0·988 for 12 months, and AUCS from 0·713 to 0·973 for 24

months). The best prediction was for genitourinary, head and neck, lung, skin and upper gastrointestinal tumours.

Conclusion: Machine learning applied to information from a disease specific (cancer) database and the EAR can be used to predict clinical outcomes. Importantly, the approach described made use of a digital data that is already routinely collected but under-exploited by clinical health systems.

Strengths and limitations of this study

- This is the first study using machine learning of both administrative and registry data for cancer survival prediction.
- A single prognosis model is produced across all cancers, improving prediction accuracy on rare cancers.
- This is a retrospective study in a single centre.

Introduction

Over the past two decades there has been an explosion in the use of digital footprints to monitor and predict human behaviours. The source of data used for this purpose is our on-line use of the internet, the emails we send and transactions we make. Analysis of these footprints through machine learning techniques (MLT) have been exploited in the public domain by government and business to predict behaviours and inform investment decisions. In research MLT have also been used to analyse gene expression data, [1, 2] and for medical image analysis. [3, 4] However to date, there has been little exploration of these methodologies in the clinical setting. We hypothesised that MLT may offer a paradigm shift in clinical medicine that can address core issues with large and complex datasets. These techniques offer the potential to derive adaptive systems from diverse datasets, discover latent connections between data items, and to predict outcomes.

Most hospitals routinely collect large digital electronic administrative records (EAR). These are primarily used for organisational financial management. Historically, they have not been used extensively for clinical or research purposes. If these large data sets are able to be exploited using MLT it may open the way to optimise the use of collected administrative data to assist in predicting patients outcome, planning individualised patient care, monitoring resource utilisation, and improving institutional performance. [5, 6] The accurate assessment of comorbid status would improve assessment of prognosis and guide treatment decisions. [7-10] Other important information that may be contained or inferred

from an EAR includes geographical and demographic data, socioeconomic status, and history of health care facility utilisation. [2, 11, 12]

In this study, using cancer outcome prediction as a model, we wished to test the hypothesis that routinely collected digital health data, if analysed by state of the art, validated, machine learning techniques could be used to assist conventional tools in predicting clinical outcomes.

Accurate prediction of survival in patients with cancer remains a challenge due to the ever-increasing heterogeneity and complexity of cancer, treatment options, and patient populations. If achieved, reliable predictions could assist personalised care and treatment, and improve institutional performance in cancer management. In current practice clinicians use data collected at the bedside in consultations, medical records or purpose built cancer registries to aid prognostication and decision making.

The notion of using MLT to predict cancer prognosis from clinical and pathological data is not a new one. [13, 14] However, with the advent of more sophisticated and better validated techniques, not only is more accurate prediction possible, but the range of data incorporated into decision aids can be increased. [15-17]. The need to improve cancer care systems by creating linkages between registries and epidemiological surveillance through analysis of complex and large clinical databases has recently been highlighted. [18, 19]

In this study we tested the capability of MLT to predict patient outcomes in a heterogeneous cohort of cancer patients. We have interrogated two data sets: the first a

purpose-built cancer specific registry (ECO) containing demographic and tumour-related data items according to an Australian nationally agreed protocol; the second a hospital digital data set containing information about the patient's previous admissions and presentations (EAR). Finally, in a test group of 94 patients, we examined the performance of machine-learning methods in aiding a panel of expert clinicians in predicting patient survival.

Patients and Methods

Study design

This is a retrospective study using the electronic administrative record (EAR) and a specialised cancer registry (ECO) from Barwon Health, the only public tertiary institution in a region of Australia with more than 350,000 residents. With a unified hospital identity number in use across the region, Barwon Health's EAR provides a single point of access for information on patient encounters with the health system, including hospitalizations, ED visits, medications, and treatments. In addition, the Andrew Love Cancer Centre at Barwon health has a specialised cancer registry called ECO, which captures clinical data for patients in the region. ECO records information on demographics, primary tumour and metastatic tumour, cancer stage, tumour size, lymph nodes, and breast tumour specific information. Treatment type, outcomes, including death, and recurrence information (primary and metastatic) are also recorded. Table 1 shows the variables used for survival prediction. The cohort for this study consists of 963 patients identified in ECO who were first diagnosed in year 2009. The study completion date was October 31, 2012; therefore all patients had at least 2 year and 10 months follow-up. Among these patients, 736 patients also had records in the EAR. Ethics approval was obtained from the Hospital and Research Ethics Committee at Barwon Health (number 12/83). Deakin University has reciprocal ethics authorization with Barwon Health.

Analyses

The analyses centred on predicting cancer survival since the date of diagnosis, defined as the date of tumour resection. Each patient was a unit of observation in the predictive problem: Patient data collected prior to the diagnosis date were used to construct the independent variables; Survival status in a period following the assessment was the dependent variable. Two analyses were performed: The first compared survival prediction made by machine learning models and the clinician panel, based on only information from ECO. The second analysis evaluated the added discriminative power provided by EAR, by comparing the best machine learning models using three sets of predicting variables: variables from ECO (Table 1), variables from EAR (appendix), and the union of the two.

Although a survival analysis model (e.g., a proportional hazards model [20]) is commonly used in modelling risk factors, such models are not designed to predict events. In this study, survival was directly modelled using classification models to optimize prediction accuracy.

Comparing predictions by machine learning models and clinician

In the first analysis, all 963 patients in the ECO registry were randomly divided into a derivation cohort of 869 patients and a validation cohort of 94 patients (Table 2). To collect clinician prediction, patients in the validation cohort were assigned to a panel of five oncologists for survival prediction. For each patient, the oncologist was asked to estimate the survival probabilities based on the independent variables in Table 1. All clinicians estimated the patient’s survival status by producing a probability for each of the three time periods—6 months, 1 year, and 2 years. When making this assessment the clinicians did not have knowledge of the treatment type offered or given to the patient. Three machine-learning models were trained on the derivation cohort using the same set of independent variables, one for each prediction period. Each of the machine learning models was an ensemble of 400 support vector machines [21] with linear kernel (i.e., the output of the model was the average of 400 support-vector-machine outputs in Platt's a posteriori probabilities[22]). Ensemble was used to control the variability introduced by $L1$ feature selection. Each of the support vector machines was trained using a random 80% subsampling (without replacement) of the derivation cohort.[23] The soft margin parameter (C) of SVM was selected through cross-validation. Two measures were taken to improve the training process. First, to compensate for the imbalance between the two outcomes (there were more survivals than deaths), we oversampled the non-surviving cases by 50% in each training subsample. Next, variable selection was performed through fitting a generalized linear model with elastic net regularization [24] (alpha parameter set to 0·1 and lambda parameter selected using 5-fold internal cross-validation) and variables with zero coefficients were removed. After the machine learning models were constructed, they were applied to predict survival probabilities for each patient in the validation cohort. Both the clinician and model predictions were validated with the actual outcomes in the ECO registry. Prediction performance was measured using the area under the ROC curve (AUC), also known as the C-statistic. [25] 95% confidence intervals of AUCs were computed using 1000 bootstrap samples of validation cohort.

Comparing discriminative information from specialized registry and routine data

The second analysis compared the discriminative power of two data sources (ECO and EAR). In this analysis, clinician predictions were not solicited. Among the 869 patients in the derivation subset of Cohort 1, only 664 have records in the EAR and these patients were included in the second analysis (Cohort 2, Table 2). Survival prediction models were derived based on three sets of independent variables: 1) independent variables from EAR (EAR only); 2) independent variables from ECO (ECO only); 3) the union of the two sets (*EAR + ECO*). Similar to the previous analysis, the models were trained using 400 random subsamples comprising 80% data of the cohort-2 and the modelling process was identical. However, the models were evaluated not using the validation cohort. Instead, for each 80% subsample, the remaining 20% was used to compute the AUC and its 95% confidence interval.

The Wilcoxon rank-sum test was applied to answer the following comparison problems:

1. Does *ECO only* provide more discriminative power than *EAR only*?
2. Does *EAR + ECO* provide more discriminative power than *EAR only*?
3. Does *EAR + ECO* provide more discriminative power than *ECO only*?

Details of the machine learning model and the predictor variables can be found in the Appendix.

Results

The cohorts for the two analyses are summarized in Table 2. The comparison between the algorithmic predictions and the clinician predictions are summarized in Table 3. The model had comparable performance to that of the clinicians, with the performance of the machine learning model marginally better (AUC ranging from 0·76 to 0·87) than that of the clinicians (AUC ranging from 0·75 to 0·79) for all three prediction periods. This similarity in accuracy between algorithmic predictions and the clinician predictions was observed across different cancer types. Consider the predictions for six-month survival. Out of 15 breast cancer cases, the clinicians made 15 correct predictions and the algorithm made 14; Out of 18 lung cancer cases, the clinicians made 13 correct predictions and the algorithm made 14; Out of 7 haematological cases, both the clinicians and the algorithm made all predictions correctly. Similar results were observed on 12-month and 24-month survival predictions for different cancers.

Prediction of 6-month survival using the three models is shown in Table 4. There were no deaths from breast cancer during this period. Comparing the ECO model with the EAR model, AUCs were comparable for colorectal, genitourinary, haematological, head and neck and skin tumours. The EAR model was significantly better ($p < 0·05$) for rare tumours; CNS, upper gastrointestinal and unassigned primary source tumours. For each tumour type, the model using both ECO and EAR data yielded similar or better performance to the models using information from only one of the two databases. AUCs for the combined model ranged from 0·76 to 1·0. The combined data model showed particularly improved performance over ECO data (p value <·05) for all tumour streams except and breast and CNS tumours.

Data for 12-month survival prediction is shown in Table 5. Cancer-specific ECO data yielded better prediction than EAR data (p < 0·05) for gynaecological, haematological lung, skin and unknown primary cancers. Otherwise, ECO and EAR models yielded generally similar results. The model using combined data performed better than EAR (p value < ·05) for all tumour streams other than CNS, head and neck and upper gastro tumours. The model using combined data was better than (P < 0·05) ECO for all cancers except breast, CNS, gynaecological and haematological cancers.

Table 6 shows data for 24-month survival prediction by the three models. The ECO model yielded superior prediction (p value < ·05) to the EAR model for breast, genitourinary, gynaecological, lung, skin and unknown primary cancers, while the EAR model was superior to the ECO model for haematological and head and neck tumours. Once more the model that performed the best was that derived from both ECO and EAR data with AUCs ranging from 0·71 to 0·97 across the range of cancers and particularly enhanced performance for all cancers except breast, colorectal, gynaecological and unknown primary tumours compared to the ECO. In summary, over all time periods, the performance of the combined model was better than ECO (p < 0·05) for genitourinary, head and neck, lung, skin and upper gastrointestinal tumours.

One of the key advantage of using machine learning technique is that it can combine the large number of non-clinical factors with the few clinical risk factors. In this study, the model selected most of the known clinical risk factors including *patient age*, *cancer staging*, *performance status,* and *tumour size*. In addition it also found some useful non-clinical risk factors, including the type of the last hospital admission (emergency vs. elective), the frequency of ED visits within the previous 3 months and 6 months (related to both cancer and other medical conditions).

Discussion

In this study, using cancer outcome prediction as a model, we wished to test the hypothesis that routinely collected digital health data, if analysed by MLT, could be used to assist conventional tools in predicting clinical outcomes.

Applying machine learning to data from the electronic administrative record (EAR) alone predicted clinical outcomes with reasonable accuracy. Using the purpose built ECO data set, the predictive tool also performed well across a broad range of cancer types, and in both cases the predictive accuracies were at least as good as that of a panel of five expert clinicians. Importantly a predictive tool derived from both the purpose built clinical registry and administrative data had even greater predictive ability.

The wealth of administrative data contained in the EAR includes information on comorbid conditions and previous clinic and hospital attendances as well as a drug history. There is considerable potential to use this data to improve clinical care across a spectrum of diseases. [5, 6]

Most patients in the study were followed up for 3 years, which may not be adequate to capture all oncologic outcomes, especially for those cancers with low mortality rate. We have designed this study as retrospective and in a single centre; it will be of major interest to observe how it performs in a variety of settings. The number of cases used to assess performance of the models is relatively small. The strengths include the comparison of machine learning tools with expert clinical opinion and the fact that very detailed and well-validated data was available both directly related to the cancer and that contained in the EAR. The generic nature of this approach makes it unnecessary to generate separate predictive models for different types of cancer. This was a particular advantage for rarer forms of cancer where predications using more conventional methods are very challenging.

Predictive tools derived from clinical data items have considerable potential to improve clinical care, but must be suitably optimised and shown to perform equally well in diverse clinical settings. [26, 27] Clinical databases have become more widely available and increasingly complex in recent years. The extent and complexity of data available to clinicians means that novel approaches to managing data and supporting clinical decisions are needed. Machine learning approaches can not only cope with complex datasets, but also adapt in real time and across different clinical settings.

The approach used in this study offers superior performance to previous machine learning approaches to predicting cancer survival. [13-17] Previous models have been derived for single cancer types, or for a limited range of cancers. The model described here performed well across a wide range of cancers. One advantage of this generic approach may be the ability to predict outcomes in less common cancers where limited data might preclude development of specific models. The fact that our model derived from administrative and cancer-related data performed slightly better than a panel of expert clinicians not only validates the potential utility of the model but suggests that it may be useful in assessing quality of care and also in settings where specialist care is not available. An alternative approach to borrow information across different cancer types is call multi-task learning. We are currently exploring this approach as well.

Clinical outcomes in any illness are determined not only by specific factors related to the illness itself but also by the patient's general state of health and by the presence of other chronic medical conditions often coded in an EAR if the individual traffics the health service.[7-10] As well, a particularly novel and important aspect of the use historical data from the EAR in machine learning is that it effectively captures the health care institutions current and previous performance. These data can be applied to any individual entering the system with a newly diagnosed cancer, as we have modelled here. As well they could also be used for quality and performance monitoring.

In conclusion, machine learning applied to information from a disease specific (cancer) database and the EAR can be used to predict outcomes. Improved prediction of outcome has the potential to help clinicians make more meaningful decisions about treatment and to assist with planning of future social and care needs. Most importantly, the approach described makes use of digital data that is already routinely collected but under-exploited by clinical health systems.

References

1. Zhao X, Rodland EA, Sorlie T, Naume B, Langerod A, Frigessi A, Kristensen VN, Borresen-Dale AL, Lingjaerde OC. Combining gene signatures improves prediction of breast cancer survival. *PLoS ONE [Electronic Resource]*. 2011;6:e17845
2. Chang CM, Su YC, Lai NS, Huang KY, Chien SH, Chang YH, Lian WC, Hsu TW, Lee CC. The combined effect of individual and neighborhood socioeconomic status on cancer survival rates. *PLoS ONE [Electronic Resource]*. 2012;7:e44325
3. Li C, Zhang S, Zhang H, Pang L, Lam K, Hui C, Zhang S. Using the k-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. *Computational & Mathematical Methods in Medicine*. 2012;2012:876545
4. Huang ML, Hung YH, Lee WM, Li RK, Wang TH. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *Journal of Medical Systems*. 2012;36:407-414
5. Appari A, Eric Johnson M, Anthony DL. Meaningful use of electronic health record systems and process quality of care: Evidence from a panel data analysis of u.S. Acute-care hospitals. *Health Services Research*. 2013;48:354-375
6. Fitzhenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, Reeves RM, Aronsky D, Elkin PL, Messina VP, Speroff T. Exploring the frontier of electronic health record surveillance: The case of postoperative complications. *Medical Care*. 2013;51:509-516
7. Lund L, Borre M, Jacobsen J, Sorensen HT, Norgaard M. Impact of comorbidity on survival of danish prostate cancer patients, 1995-2006: A population-based cohort study. *Urology*. 2008;72:1258-1262
8. Tetsche MS, Norgaard M, Jacobsen J, Wogelius P, Sorensen HT. Comorbidity and ovarian cancer survival in denmark, 1995-2005: A population-based cohort study. *International Journal of Gynecological Cancer*. 2008;18:421-427
9. Lieffers JR, Baracos VE, Winget M, Fassbender K. A comparison of charlson and elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data. *Cancer*. 2011;117:1957-1965
10. Braithwaite D, Moore DH, Satariano WA, Kwan ML, Hiatt RA, Kroenke C, Caan BJ. Prognostic impact of comorbidity among long-term breast cancer survivors: Results from the lace study. *Cancer Epidemiology, Biomarkers & Prevention*. 2012;21:1115-1125
11. Jones LE, Doebbeling CC. Beyond the traditional prognostic indicators: The impact of primary care utilization on cancer survival. *Journal of Clinical Oncology*. 2007;25:5793-5799
12. Sant M, Minicozzi P, Allemani C, Cirilli C, Federico M, Capocaccia R, Budroni M, Candela P, Crocetti E, Falcini F, Ferretti S, Fusco M, Giacomin A, La Rosa F, Mangone L, Natali M, Leon MP, Traina A, Tumino R, Zambon P. Regional inequalities in cancer care persist in italy and can influence survival. *Cancer Epidemiology*. 2012;36:541-547
13. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, Jr., Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*. 1997;79:857-862
14. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. *Oncology*. 1999;57:281-286

15. Manilich EA, Kiran RP, Radivoyevitch T, Lavery I, Fazio VW, Remzi FH. A novel data-driven prognostic model for staging of colorectal cancer. *Journal of the American College of Surgeons*. 2011;213:579-588, 588.e571-572

16. Gao P, Zhou X, Wang ZN, Song YX, Tong LL, Xu YY, Yue ZY, Xu HM. Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. The tnm staging system. *PLoS ONE [Electronic Resource]*. 2012;7:e42015

17. Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW. Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer*. 2012;15:230-238

18. Johnson CJ, Weir HK, Fink AK, German RR, Finch JL, Rycroft RK, Yin D, Accuracy of Cancer Mortality Study G. The impact of national death index linkages on population-based cancer survival rates in the united states. *Cancer Epidemiology*. 2013;37:20-28

19. Khoury MJ, Lam TK, Ioannidis JP, Hartge P, Spitz MR, Buring JE, Chanock SJ, Croyle RT, Goddard KA, Ginsburg GS, Herceg Z, Hiatt RA, Hoover RN, Hunter DJ, Kramer BS, Lauer MS, Meyerhardt JA, Olopade OI, Palmer JR, Sellers TA, Seminara D, Ransohoff DF, Rebbeck TR, Tourassi G, Winn DM, Zauber A, Schully SD. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology, Biomarkers & Prevention*. 2013;22:508-516

20. Cox DR, Oakes D. *Analysis of survival data*. CRC Press; 1984.

21. Cortes C, Vapnik V. Support vector machine. *Machine learning*. 1995;20:273-297

22. Lin H-T, Lin C-J, Weng RC. A note on platt's probabilistic outputs for support vector machines. *Mach. Learn.* 2007;68:267-276

23. Politis D, Romano J, Wolf M. *Subsampling*. Springer-Verlag, New York; 1999.

24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 2010;33:1

25. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*. 2005;27:83-85

26. Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. *BMC Medical Research Methodology*. 2012;12:102

27. Chen HC, Chen JJ. Assessment of reproducibility of cancer survival risk predictions across medical centers. *BMC Medical Research Methodology*. 2013;13:25

Table 1: ECO variables used for survival prediction

patient demographics
post code
Gender
Age
tumour characteristics
primary site (in ICD-10 code)
tumour stream
morphology (in ICD-O-3 code)
histologic grade
metastatic sites
most valid basis of diagnosis
performance status diagnosis
stage basis (pathological or clinical)
stage (TNM)
tumour size
nodes taken
positive nodes
breast cancer related variables
oestrogen receptor
progesterone receptor
human epidermal growth factor receptor 2 (HER2)

Table 2: Characteristics of Derivation and Validation Cohorts

	Cohort 1: ECO		Cohort 2: ECO and EAR (n=664)
	Derivation (n=869)	Validation (n=94)	
Age (SD)	67·6 (14·6)	68·4 (13·6)	66·3(14·9)
Gender: Males	487 *	48	381
Tumour stream			
Genitourinary	172	21	135
Colorectal	140	14	115
Lung	121	18	96
Breast	122	15	74
Haematological	99	7	85
Upper gastro	83	9	57
Skin	36	1	28
Head and Neck	35	0	30
Gynaecological	19	4	17
CNS	15	1	9

Unknown primary	38	9	26

*2 unspecified

Table 3: Performance of survival prediction: comparison between machine learning method and clinicians

Survival Period	AUC (95% CI)	
	Clinician panel	Machine learning model
6 months	0·79 (0·76, 0·81)	0·87 (0·85, 0·89)
1 year	0·79 (0·76, 0·81)	0·80 (0·77, 0·82)
2 years	0·75 (0·73, 0·78)	0·76 (0·74, 0·79)

Table 4: Prediction performance of machine learning algorithms: 6-month survival

	Area Under ROC Curve (95% CI)		
Cancer type	**EAR only**	**ECO only**	**EAR + ECO**
Genitourinary	·81 (·77, ·85)	·82 (·78, ·86)	·88 (·85, ·91) [*,†]
Colorectal	·84 (·80, ·88)	·85 (·81, ·89)	·88 (·84, ·91) [*,†]
Lung	·71 (·67, ·76)	·73 (·69, ·77) [*]	·77 (·73, ·82) [*,†]
Breast		no deaths in the period	
Haematological	·73 (·68, ·79)	·74 (·69, ·79)	·76 (·71, ·81)
Upper gastro	·74 (·69, ·78)	·64 (·60, ·69)	·84 (·80, ·87) [†]
Skin	·84 (·77, ·90)	·85 (·79, ·91)	·91 (·86, ·96) [*,†]
Head and neck	·66 (·61, ·71)	·70 (·64, ·75)	·77 (·72, ·82) [*,†]
Gynaecological	·97 (·94, ·99)	·99 (·98, 1·0) [*]	1·0 (·99, 1·0) [*]
CNS	·89 (·85, ·94)	·84 (·78, 0·90)	·82 (·77, ·88)
Unknown primary	·92 (·89, 95)	·79 (·75, ·84)	·90 (·87, ·93) [*,†]

[*] Significantly greater than *EAR only*. [†] Significantly greater than *ECO only*.

Table 5: Prediction performance of machine learning algorithms: 12-month survival

	Area Under ROC Curve (95% CI)		
Cancer type	***EAR only***	***ECO only***	***EAR + ECO***
Genitourinary	·79 (·75, ·83)	·79 (·75, ·83)	·84 (·80, ·87) [*,†]
Colorectal	·82 (·78, ·86)	·83 (·79, ·86)	·87 (·83, ·90) [*,†]
Lung	·73 (·69, ·77)	·78 (·73, ·82) [*]	·82 (·78, ·86) [*,†]
Breast	·71 (·65, ·78)	·90 (·86, ·94)	·92 (·89, ·96) [*]
Haematological	·63 (·59, ·68)	·70 (·66, ·75) [*]	·69 (·64, ·74) [*]
Upper gastro	·62 (·57, ·66)	·70 (·65, ·74) [*]	·72 (·68, ·76) [*]
Skin	·76 (·71, ·88)	·89 (·85, ·93) [*]	·93 (·90, ·96) [*]
Head and neck	·77 (·73, ·88)	·68 (·63, 73)	·79 (·75, ·84) [†]
Gynaecological	·95 (·92, ·97)	1·0 (1·0, 1·0) [*]	·99 (·98, 1·0) [*]
CNS	·66 (·58, ·73)	·68 (·61, ·76)	·69 (·63, ·76)
Unknown primary	·87 (·84, ·91)	·81 (·77, ·85)	·88 (·84, ·91)

[*] Significantly greater than *EAR only*. [†] Significantly greater than *ECO only*.

Table 6: Prediction performance of machine learning algorithms: 24-month survival

	Area Under ROC Curve (AUC)		
Cancer type	**EAR only**	**ECO only**	**EAR + ECO**
Genitourinary	·73 (·69, ·78)	·84 (·81, ·88) [*]	·86 (·82, ·89) [*,†]
Colorectal	·76 (·72, 80)	·76 (·72, ·80)	·76 (·72, ·80)
Lung	·74 (·69, ·78)	·78 (·73, ·82) [*]	·82 (·79, ·86) [*,†]
Breast	·67 (·61, ·73)	·86 (·82, ·90) [*]	·88 (·84, ·92) [*]
Haematological	·73 (·68, ·77)	·70 (·66, ·75)	·80 (·76, ·84) [*,†]
Upper gastro	·81 (·77, ·85)	·77 (·72, ·81)	·87 (·83, ·90) [*,†]
Skin	·71 (·65, ·76)	·85 (·80, ·89) [*]	·94 (·92, ·97) [*,†]
Head and neck	·74 (·70, ·78)	·66 (·51, ·61)	·71 (·67, ·76) [†]
Gynaecological	·96 (·94, ·99)	·99 (·98, 1·0) [*]	·97 (·95, ·99)
CNS	·83 (·78, ·89)	·87 (·82, ·93)	·96 (·93, ·99) [*,†]
Unknown primary	·74 (·70, ·79)	·78 (·74, ·82) [*]	·80 (·76, ·84) [*]

[*] Significantly greater than *EAR only*. [†] Significantly greater than *ECO only*.

Appendix

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

Derivation of the machine learning model

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

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

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

Predictors for the machine learning models

Table 1 EMR-based predictors

demographics

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

cancer specific diagnoses

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

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

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

Table 2 ECO-based predictors.

patient demographics

Gender
Age

tumour characteristics

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

breast cancer related variables

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

References

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

STARD checklist for reporting of studies of diagnostic accuracy

(version January 2003)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	1-4
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	5-8
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	9
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	9
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	Yes
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	retrospective
Test methods	7	The reference standard and its rationale.	9-10
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	9-10
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	9-10
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	N/A
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	N/A
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	10-12
	13	Methods for calculating test reproducibility, if done.	10-12
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	9
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	21
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	N/A
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	N/A
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	N/A
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	N/A
	20	Any adverse events from performing the index tests or the reference standard.	N/A
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	23-25
	22	How indeterminate results, missing data and outliers of the index tests were handled.	N/A
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	N/A
	24	Estimates of test reproducibility, if done.	N/A
DISCUSSION	25	Discuss the clinical applicability of the study findings.	15-17