Chapter 1

A Preoperative Molecular Prognostic for Pancreas Cancer

Thesis: A preoperative prognostic tool for pancreas cancer can be developed to discriminate good between and poor prognosis patients more reliably than current methods.

Summary For those patients fortunate enough to be diagnosed with a resectable tumour, surgical removal of the primary cancer is the best first-line therapy for pancreas cancer. However, the significant morbidity associated with pancreas cancer resection makes it cruicially important to only operate on the patients who will derive a net benefit from the procedure. Identifying just those patients who will respond to resection remains a serious challenge in pancreas cancer treatment: current criteria to select patients for resection perform poorly, and consequently many patients undergo a complex procedure, with serious effects on future quality of life, for little benefit. Tumour biomarkers have the potential to dramatically refine morphology-based staging criteria by supplying a direct readout of tumour biology, and recent technological developments have enabled the preoperative measurement of tissue biomarkers in pancreas cancer. The ability to measure pancreas cancer tissue biomarker levels preoperatively, combined with the enhanced information on disease state available from tissue biomarkers, finally enables the development of preoperative staging systems that accurately identify pancreas cancer patients for resection. This chapter details the development and validation of the Pancreas Cancer Outcome Predictor (PCOP), a two-biomarker prognostic tool for resectable pancreas cancer, that is in principle pre-operatively assessable, and can assist in making personalised treatment decisions.

1.1 Introduction

For patients with a resectable tumour and no known metastases, surgical removal of the primary tumour is the current recommended first-line therapy for pancreas cancer, and the only intervention offering the realistic possibility of a cure [?]. However, pancreas cancer resection is a major procedure, with the potential for serious complications, morbidity, and reduced quality of life following recovery [?]. Due to the substantial negative effects of surgery, the decision of whether or not to perform curative-intent resection should balance the risks of surgery against its expected benefits, for each individual case.

Unfortunately, current practice guidelines recommend that curative-intent surgery be offered to all metastasis-free patients with a resectable tumour, with no consideration of personal benefit [?]. This blanket approach to selecting patients for curative resection has proven to be highly inadequate. Even following pathologically complete tumour removal and adjuvant chemotherapy, more than 70% of current pancreas ductal carcinoma patients will relapse with, and ultimately succumb to, distant metastases [?]. These occult metastases must have been present prior to removal of the primary tumour, yet were undetectable during initial investigations, and their presence means that any curative-intent resection was futile. As a result, the majority of 'curative' resections that are undertaken based on current selection criteria are performed on patients with occult metastases, have no hope of actually effecting a cure, and would not have been undertaken at all if the presence of metastatic disease had been known prior to surgery. Better methods to select patients for resection are urgently needed.

The decision of whether or not to resect will always involve close consultation between the patient and doctors, comparing the costs and benefits of surgery as appropriate to each patient's case. The downsides of resection are well-understood, but the benefit to be gained is challenging to quantify and communicate, being highly dependent on the many particulars of an individual patient's disease. A simple approach to represent benefit from pancreas cancer resection involves survival curves, which plot the probability that a patient will be alive, at a range of times following resection (TODO example fig). A survival curve distills the information from an arbitrary number of prognostic factors into a single simple figure, to provide an intuitive overview of a patient's expected disease course. If accurate, such curves can provide a concrete measure of benefit from resection, and thus provide invaluable input into the treatment decision process. Survival curves can be involved to calculate, which has likely limited their historical use. However, the modern prevalence of computers removes this barrier, and even a modest device such as a smart phone can easily run prognostic tools capable of generating accurate personalised patient survival curves.

A number of pancreas cancer grading and schemes and prognostic tools have been described, but inconsistent performance, or a reliance on information that can only be known post-operatively, limits their use in pre-operative decisions. The level of serum carbohydrate antigen 19-9 (CA-19-9) is a well-characterised biomarker of pancreas cancer, with high levels correlating with increased tumour burden, lower probability of resectability, increased post-resection recurrence, and worse prognosis [?, ?, ?, ?]. CA-19-9 levels are easily determined pre-operatively, but the use of this marker is complicated by a lack of consensus on threshold concentrations, the elevation of CA-19-9 levels by a number of conditions other than pancreas cancer, and the complete absence of this marker in approximately 10% of the general population [?]. Additionally, although CA-19-9 levels are statistically associated with post-resection recurrence by distant metastasis, a very low positive predictive value (PPV) renders the biomarker unhelpful when deciding whether or not to resect [?].

The current standard prognostic tool for pancreas cancer is the Memorial Sloan-Kettering Cancer Center (MSKCC) nomogram [?], which integrates a number of clinico-pathological variables (CPVs) to arrive at point estimates of survival post-resection. Unfortunately, its clinical utility is small: as it relies on information that is only available following resection, the MSKCC nomogram is only useful in a post-operative context, and cannot assist in pre-operative decisions to resect. This severely limiting reliance on postoperative variables is made necessary by the fact that all strong classical prognostic factors in pancreas cancer (such as lymph node infiltration, resection margin status, or histological grade [?]) can only be reliably measured following resection. Any prognostic tool for pancreas cancer that relies heavily on classical CPVs will very likely share this same reliance on post-operative variables, and so an effective pre-operatively assessable prognostic will need to shirk classical CPVs, and leverage novel pre-operative measures of prognosis.

Levels of tissue biomarkers directly reflect cellular state, and thus have the potential to predict cancer behaviour far more reliably than macroscopic CPVs. Given that most pancreas cancer patients who undergo curative resection quickly recur due to occult metastases, biomarkers of metastasis have the potential to identify those patients who are likely to already have occult metastatic disease at the time of surgery, and thus better inform the decision to resect. Two such biomarkers of metastasis are the cancer cell levels of the epithelial to mesenchymal transition (EMT)-related S100A2, and S100A4 proteins, both of which are strongly predictive of outcome following resection, and appear to reflect the presence of a pro-metastatic invasive phenotype in the cancer [?, ?, ?]. Despite this promise, these tissue biomarkers have to date only been assessed in bulk tissue samples collected during surgery, and their utility, or even measurability, in a pre-operative setting, is untested.

Recent techological developments have made possible the pre-operative measurement of tissue biomarkers during endoscopic ultrasound (EUS), a routine diagnostic modality for pancreas cancer. Immunohistochemical (IHC) staining has been successfully performed on fine needle aspirate (FNA) biopsies of pancreas neoplasms collected during EUS [?, ?, ?], and in principle EUS-FNA-IHC could form the basis of a routine pre-operative biomarker measurement methodology in pancreas cancer. This proposed biomarker measurement approach utilises only techniques that are commonly available in pancreas cancer treatment centres, and thus has the potential to be rapidly integrated into current diagnostic workflows, should biomarker measurements prove to be clinically valuable.

The nexus of known biomarkers of metastatic behaviour, new pre-operatively applicable techniques to measure these biomarkers, and multiple large, clinically annotated cohorts of resected pancreas cancer, presents an opportunity to address the pressing need for better criteria to select patients for pancreas cancer resection. As part of the Australian Pancreatic Cancer Genome Initiative (APGI), as well as other work, the group has collected tissue measurements of S100A2 and S100A4 biomarkers, and detailed patient follow-up, for a large number of cases of pancreas cancer from a range of independent cohorts. These cases will be used to develop the Pancreas Cancer Outcome Predictor (PCOP), a tool to predict outcome following resection, using tissue levels of S100A2 and S100A4 as major prognostic factors. This initial version of PCOP is based on biomarker measurements made on tissue collected during resection, and thus will not be directly applicable pre-operatively. However, pilot study data will be used to demonstrate that levels of S100A2 and S100A4 measured by pre-operative EUS-FNA-IHC correlate well to tissue levels of the biomarkers measured on operative specimens, indicating that a more refined version of PCOP trained on pre-operative data will be equally effective.

The majority of pancreas cancer resection procedures today are performed on patients who should never have been offered surgical resection at all. These patients have undetected metastases at the time of surgery, and will derive little benefit from a major operation, that has serious impacts on quality of life. Current tools for patient staging and estimation of prognosis are either ineffective at identifying patients at risk for occult metastases, or only applicable post-operatively, and so cannot be used to inform the decision of whether or not to resect. Tissue biomarkers of metastatic potential might identify, pre-operatively, those patients who have a high likelihood of metastatic disease, greatly assisting disease management decisions. This metastasis prediction can be integrated with other clinical variables to yield personalised estimates of prognosis over time, that are well-suited to . This chapter describes the use of pre-operatively assessable variables, including biomarker measurements, to create PCOP, a tool that produces estimates of prognosis. PCOP provides a natural way to show the influence of risk factors on a patient's personalised prognostic path, and thus can assist in making treatment decisions appropriate for each individual pancreas cancer patient.

1.2 Results

Data from the large, retrospectively-acquired New South Wales Pancreatic Cancer Network (NSWPCN) cohort were used to derive PCOP, a tool to predict the survival of pancreas cancer patients following curative-intent resection. Discrimination and calibration of PCOP were verified on two independent surgical cohorts. Data from an EUS-FNA-IHC pilot study established that pre-operatively assessed tissue biomarker levels reflected measurements from operative biopsies, and therefore that PCOP could be translated to a pre-operative decision setting.

Prognostic variables and biomarkers

As the aim was to develop a prognostic predictor that could be applied pre-operatively, only factors that could be practically measured pre-operatively were considered for inclusion in the PCOP. The traditional CPVs that were judged to be pre-operatively assessable were patient sex, patient age at diagnosis, tumour location (dichotomised as head of pancreas vs other location), and size of the tumour's longest pathological axis. In addition to these traditional factors, the dichotomised tissue levels of S100A2 and S100A4 proteins were included as candidate biomarkers in the construction of the PCOP. Pre-operative blood levels of the biomarker CA-19-9 were available for a subset of the training cohort, but none of the validation sets; for this reason, and the marker's generally poor performance in isolation [?], CA-19-9 levels were not considered for inclusion in the PCOP.

Pre-operative measurements of tumour size (for example, by computed tomography (CT) X-ray or EUS) were not available in the training and validation sets, and were approximated by post-operative measurements during the development and testing of this nomogram. Similarily, biomarker measurements were approximated using IHC staining of tissue collected during resection, as only very limited pre-operative EUS-FNA samples were available in the cohorts used. The implications of these approximations for the prognostic tool developed here, as well as for future work, are considered in the discussion.

[1]

Cohorts and Characteristics

General characteristics of the NSWPCN, Glasgow, and Dresden cohorts are summarised in Table 1.1. Reliable data on adjuvant and neoadjuvant chemotherapy treatment was not available in all cohorts, and may form a confounding

[1] MP Fatal: For the disc: Although the correlation between CT and EUS estimates of tumour size, and actual size upon resection, is respectable [?], full clinical validation of this prognostic's use in a pre-operative setting will require ...

factor. The NSWPCN training cohort contained a small subgroup of patients with abnormally long recorded survival times (> 3000 days, 7/256 patients), that were strongly suspected to represent data errors, either as a consequence of incorrect coding following loss to follow-up, or misdiagnosis. Given the age of the cohort, it was deemed impractical to revisit the original records to check these patients, and so all patients with recorded survival times exceeding 3000 days were excluded from the NSWPCN training data. The NSWPCN cohort characteristics in Table 1.1 have been calculated on the 249 patients that passed the 3000 day data quality cutoff.

The four cohorts had broadly similar marginal survival functions (Figure 1.1), although these were statistically distinct (logrank $P = 5.7 \times 10^{-6}$). In particular, the NSWPCN cohort had the lowest survival rate of all cohorts from one year following resection; the better outcome of the more modern cohorts over NSWPCN may be a consequence of differing adjuvant chemotherapy rates between cohorts, or reflect recent improvements in disease management that have yielded slightly improved overall survival. After correcting for all available covariates, cohort still had a significant effect on survival (likelihood ratio test $P = 3.8 \times 10^{-8}$), although there was no sigificant difference in baseline survival function between cohorts (Grambsch-Thernau test [**?**] Holm-corrected $P > 0.23$, 24 tests). Despite there being no significant indication of differences in baseline hazard between cohorts, the presence of a strong and significant cohort effect that independent of all measured variables limits the maximum possible validation performance of any prognostic predictor on these data.

There were significant differences between the cohorts in the distribution of both CPVs and biomarker scores. In particular, large variation was present in the fraction of patients with clear resection margins (range $27\% - 65\%$, Fisher exact test $P = 2.2 \times 10^{-15}$), and lymph node involvement ($66\% - 83\%$, $P = 8.3 \times 10^{-5}$), suggesting substantial variation in cohort composition. Biomarker scores were also significantly differently distributed between cohorts (S100A2 $15\% - 33\%$, $P = 1.5 \times 10^{-4}$, S100A4 $65\% - 88\%$, $P = 1.3 \times 10^{-4}$). This difference in biomarker scores is likely largely due to cohort-specific technical differences in tissue collection, processing, staining, and scoring, although cohort composition effects may also have contributed.

The large differences between training and validation cohorts provides a strong test of the ability of a prognostic tool to generalize to new cohorts, laboratory processes, and scoring pathologists. Residual unexplained effects of cohort on survival will limit the validation calibration performance attainable on these data, but clinically useful accurate discrimination of good- and poor-prognosis patients may still be achievable.

Table 1.1: Characteristics of the NSWPCN training cohort, and the APGI, Dresden, and Glasgow validation cohorts. Ordinal variables are shown as median, with quartiles in parentheses. Categorical variables for which percentages do not add up to 100% indicate the presence of minor unlisted categories.

		Training	Validation		
Characteristic		NSWPCN	APGI	Dresden	Glasgow
Number of patients		249	75	150	189
Gender	Male	49.4%	54.7%	54.7%	52.9%
Tumour location	Head	80.7%	85.3%	92.7%	100%
Excision margin status	R0	58.2%	32.0%	65.3%	27.0%
Node involvement		65.8%	78.7%	68.7%	82.5%
S100A2 positive		16.1%	14.7%	25.3%	32.8%
S100A4 positive		75.5%	65.3%	88.0%	70.9%
Disease-specific death		95.2%	68.0%	74.7%	85.2%
Size of longest axis	(mm)	30	35	35	30
		(25 - 40)	(28 - 43)	(25 - 40)	(25 - 40)
Age at diagnosis	(years)	69	67	68	64.0
		(62 - 75)	(61 - 74)	(59 - 73)	(57.8 - 69.4)
Length of follow-up	(days)	479	655	514	501
		(270 - 851)	(362 - 743)	(311 - 915	(233 - 915)

Prognostic model building and selection

Candidate prognostic models were constructed on the NSWPCN training data by iterative model fitting, evaluation, and refinement. To guard against overfitting caused by this iterative process, the NSWPCN cohort was randomly split once into model building and testing sets. All model fitting and refinement described below was performed on the 200-patient model building set, to yield three final candidate prognostic predictors. The performance of each of these three predictors was then assessed on the 49-patient model test set, and the most parsimonious high-performing model was chosen as the final prognostic predictor, for subsequent external validation.

Cohort shift The NSWPCN training cohort was collected over a long period, with patient diagnosis dates spanning the thirteen years from 1994 to 2006. Over such an extended interval, subtle changes in cohort composition or therapy may cause a shift in cohort characteristics, and reduce the prognostic performance of a model that was built on the historical data, when it is applied to contemporary cases. Cohort shift was investigated by examining the association between date of diagnosis, and all prognostic and outcome variables: in the absence of shift, no variables would be expected to change significantly over time. Date of diagnosis was not significantly associated with any other variable (7 tests, lowest $P = 0.35$); there was therefore no indication of cohort shift in the NSWPCN training data.

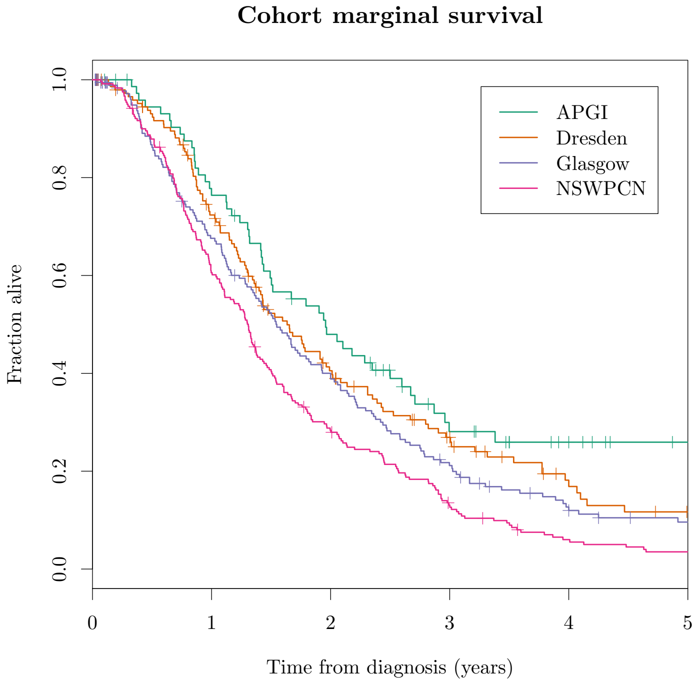
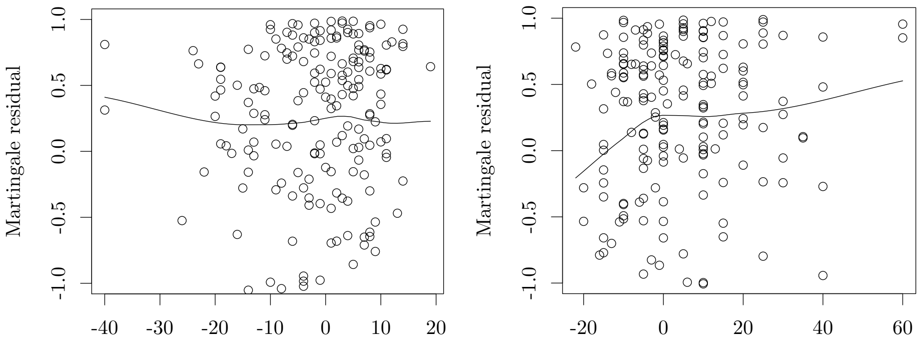


Figure 1.1: Kaplan-Meier marginal survival function estimates for the cohorts used in this chapter. Censoring times are indicated by + symbols.

Model functional form and expanded terms The Cox proportional hazard (CPH) framework was used to assess functional form for the two continuous covariates: age at diagnosis, and maximum pathological axis size. local regression (LOESS) smooths of martingale residuals [**?**] indicated a largely linear relationship for age at diagnosis (Figure 1.2(a)), and a knee-shaped form for size (Figure 1.2(b)), with the knee at approximately 0 in median-centered units. In subsequent fits this potential nonlinear size effect was modelled by adding a size$_+$ ramp term. The original set of five linear prognostic terms, plus the additional nonlinear size term, was denoted the expanded term set.

Proportional hazards assumption A Grambsch-Therneau test [**?**] on the CPH model fit using all expanded terms indicated that patient sex violated the proportional hazards (PHs) assumption ($P = 0.0104$, Figure 1.3) – in other words, the two sexes had significantly different baseline hazard shapes. To account for this effect, all subsequent models were stratified by patient sex, so that the survival of male and female patients was modelled by two different baseline hazard functions. A Grambsch-Thernau test on the stratified model indicated no further significant violations of the PH assumption (global $P = 0.4194$).



(a) Patient age (centered)

(b) Tumour size (centered)

Figure 1.2: NSWPCN prognostic predictor functional forms. Smoothed Cox model martingale residual plots indicate hazard relationships that are approximately linear for centered age (panel a), and piecewise linear for centered tumour size (panel b). For clarity, plots have been restricted to the residual range $[-1, 1]$.

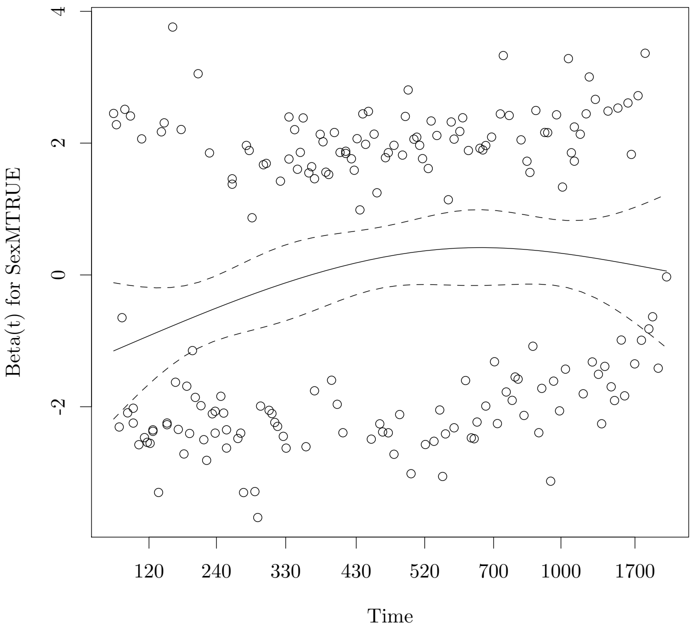


Figure 1.3: NSWPCN baseline hazard differs between patient sexes. A natural spline smooth of scaled Schoenfeld residuals for patient sex has a slope obviously differing from zero, suggesting that the baseline hazard forms differ between the two sexes, and that the combined data violates the PH assumption of Cox regression. Individual residuals are displayed as points, the natural spline smooth (df = 4) as a solid line, and approximate ±1 SE bounds as dashed lines.

Outlier removal Strongly influential or outlying samples from the full marginal Cox fit were removed from the NSWPCN building set. I considered this unusual measure to be necessary given known and unresolvable quality issues in the NSWPCN cohort data. For all subsequent work, patients with full marginal Cox model absolute deviance residuals exceeding 2.5, or any absolute DFBETAS score exceeding 0.3, were excluded from the original building set. This filter removed seven patients, reducing the size of the model building set to 193 patients.

Variable selection Stepwise variable elimination was used to select an AIC-optimal model starting from the full marginal CPH model containing all expanded terms and a sex stratum. The identified optimal CPH model used four variables: tumour location (head vs body), tumour size (linear term only), S100A2 status, and S100A4 status, in addition to the sex stratum. The final -selected set of prognostic terms (tumour location, size linear term, S100A2 binary status, S100A4 binary status, and a patient sex stratum) was denoted the reduced term set.

Model CP1 A final prognostic CPH regression model was fit to the NSWPCN model building data using only the reduced term set; this model was termed CP1. CP1 did not violate the PH assumption by the Grambsch-Therneau test (global $P = 0.794$). Predictions from model CP1 were broadly concordant with stratified Kaplan-Meier (KM) estimates across all covariate subgroups, indicating no serious lack of fit of the model (Figure 1.4).

Model GG1 Semiparametric Cox PH models such as CP1 provide a convenient framework for covariate testing and model diagnostics, but their unspecified baseline hazard term significantly complicates their use as prognostic predictors: patients can only be ranked by relative hazard, and absolute estimates of survival probabilities are unavailable. Although it is possible to approximate the baseline hazard in the Cox model, a more robust alternative is to use fully parametric models, in which the baseline hazard distribution is explicity specified. The advantages of parametric models in terms of robustness and interpretability are offset by their more stringent assumptions: if the chosen baseline distribution is unsuited to the particular data to be fit, predictions from parametric models can be very poor. Given the potential benefits of parametric models for survival prediction, a parametric alternative to model CP1 was developed, and its fit assessed. This parametric model was termed GG1.

Model GG1, employing a generalised gamma (GG) survival distribution [**?**], was fit to the NSWPCN model building data by maximum likelihood. Guided by the model functional form and baseline hazard stratification indicated by the Cox model diagnostics, the GG distribution location parameter

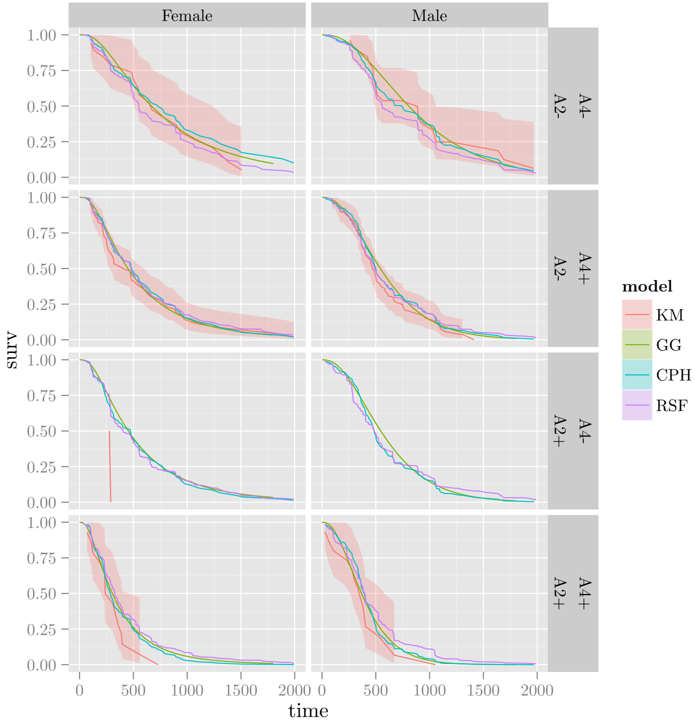


Figure 1.4: Model survival predictions agree with stratified KM estimates. KM estimates of survival probability for each combination of patient sex and biomarker status are shown as solid blue lines, with 95% confidence intervals indicated by blue ribbons. Estimates of survival probability generated by both models CP1 (red) and GG1 (green) broadly followed the form of the KM estimator, and lay within its bounds at all times, although the GG1 fit was relatively poor in some strata. Both model fitting and prediction used the NSWPCN model building set, and so these plots illustrate model goodness-of-fit, but cannot indicate possible overfitting. KM traces for the S100A2 positive, S100A4 negative group were omitted, as there were insufficient patients in this group for reliable KM estimates to be available. For all plots, tumour location, and size, were set to cohort median values.

β was made linearly dependent on all terms in the reduced set, but the shape parameters σ and λ were modelled as dependent on patient sex only. The goodness of fit of GG1 was investigated by and graphical assessment of prediction accuracy. Comparisons between GG1 predictions and KM estimates of survival indicated that the GG1 baseline distribution could not accurately model survival in some cohort subsets, particularly the S100A2 negative male group (Figure 1.4)

Model RSF Regression models like CP1 and GG1 are familiar and readily interpretable, but are heavily dependent on the analyst identifying appropriate variables and functional forms. Ensemble tree models such as random forests [**?**] naturally and automatically model nonlinearity and arbitrary level interactions, and are tolerant of large numbers of irrelevant or collinear variables, albeit at the cost of very poor interpretability, and large data and computational requirements. Random forests have been adapted to model censored data [**?**], and can provide an alternative prognostic predictor that is distinct in behaviour from CP1 and GG1, and may be able to exploit data structure not leveraged by these more classical models.

To investigate whether tree ensemble models could provide improved performance over classical approaches, a random survival forest model, termed RSF, was fit to the NSWPCN model building data. In contrast to CP1 and GG1, which used the reduced set of terms as covariates, RSF was supplied all preoperatively-assessable variables as candidate predictors.

Model selection Predictive performance of the three prognostic models (CP1, GG1, and RSF) was compared on the holdout NSWPCN model test set, to select a single high-performing parsimonious model for external validation. Performance in the interval from seven to 34 months post-diagnosis was of particular interest, as the majority of patients in the NSWPCN training set died during this period. Model GG1 was the overall best-performing model in this $7-34$ month interval, as assessed by Brier score [**?**] (Figure **??**), and all models displayed similar discriminatory ability over this period by incident/dynamic time-dependent receiver operating characteristic (TD-ROC) area under the curve (AUC) [**?**] (Figure **??**). There was no significant difference between the $7-34$ month integrated Brier score (IBS) of competing models, as estimated using 95% bootstrap confidence intervals, although all models had significantly better IBS than a marginal Kaplan-Meier survival estimator (Table 1.2). As there was no significant difference in performance between the prognostic models, the simplest model GG1 was selected to form the PCOP.

Final PCOP fit A final fit of GG1 to the full NSWPCN training data (both model building and validation patients) was made, and is summarised in Table 1.3. This fit defined the PCOP, which predicts post-resection outcome using a generalized gamma model [**?**], as

$$\begin{aligned}
T \sim GG(\beta &= 6.7446 + 0.3732[\text{Sex} = \text{Male}] - 0.2150[\text{Location} = \text{Body}] \\
&\quad - 0.0887\,\text{Size} - 0.3729[\text{S100A2} = \text{Positive}] \\
&\quad - 0.3843[\text{S100A4} = \text{Positive}], \\
\sigma &= 0.7503 - 0.2452[\text{Sex} = \text{Male}], \\
\lambda &= 0.0288 - 0.7630[\text{Sex} = \text{Male}])
\end{aligned}$$

Table 1.2: Competing models do not have significantly different IBS performance. The IBS is a combined measure of model predictive ability over a follow-up time interval, which captures both discrimination and calibration; lower numbers are better. Differences in the $7 - 34$ month IBS of competing models were calculated for each of 1,000 bootstrap samples of the NSWPCN holdout test set, and 95% BCa confidence intervals [**?**] calculated. All candidate prognostic models had significantly better IBS than the marginal KM0 model, but there was no difference between candidate models.

Comparison	Mean	95% CI
KM0 – GG1	21.1	$[2.5, 39.8]$
KM0 – CPH	20.2	$[4.5, 38.9]$
KM0 – RSF	14.5	$[5.7, 24.6]$
RSF – GG1	6.6	$[-5.6, 17.7]$
RSF – CPH	5.7	$[-2.9, 15.9]$
CPH – GG1	0.9	$[-4.1, 4.3]$

where T is an individual's failure time, GG is the generalized gamma distribution, Size is in centimetres, and [ ] is the Iverson bracket.

External validation

Discrimination

Calibration

Summary?

PCOP web application

http://54.66.150.159:3277/ [2]

1.3 Discussion

1.4 Methods

Cohort recruitment and ethics

[3]

[2] MP Fatal: Get a proper domain name?

[3] MP Fatal: For all cohorts – get on to DC?

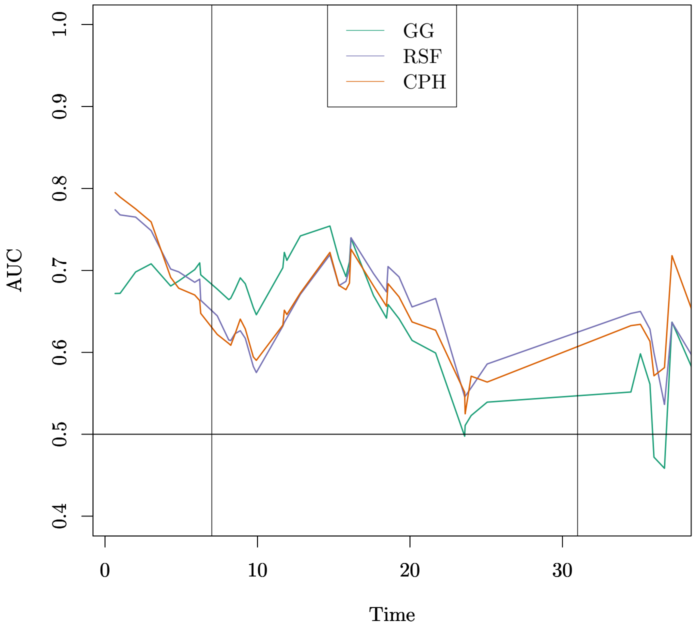


Figure 1.5: Incident/dynamic TD-ROC AUC paths for candidate models on the holdout NSWPCN model test set. Slight differences in performance were evident, with model GG providing superior discrimination up to approximately 15 months post-diagnosis, but models RSF and CPH performing better from approximately 20 months post-diagnosis. Due to the relatively small size of the holdout test set, these differences were non-significant, as assessed by pointwise bootstrap confidence intervals (confidence bands not shown).

Biomarker staining and scoring

[4]

Model building and selection

All statistical modelling was performed within the R environment. CPH and models were fit and analysed using the base package survival, and Cox model stepwise variable elimination was performed using the function stepAIC from package MASS. Generalised gamma survival models were fit using the implementation in package flexsurv[5], and package randomForestSRC supplied random survival forest functions. The random survival forest model was trained with parameters splitrule = "logrankscore", nsplit = 2, and ntree = 1000, with all other parameters set to defaults.

[4] MP Fatal: For all cohorts – get on to DC?

[5] Parameter symbols differ between the flexsurv package, and this chapter and [?]. In this chapter and [?], the generalized gamma location parameter is denoted β, and shape parameters are σ, and λ. In flexsurv, these parameters are denoted μ, σ, and Q, respectively.

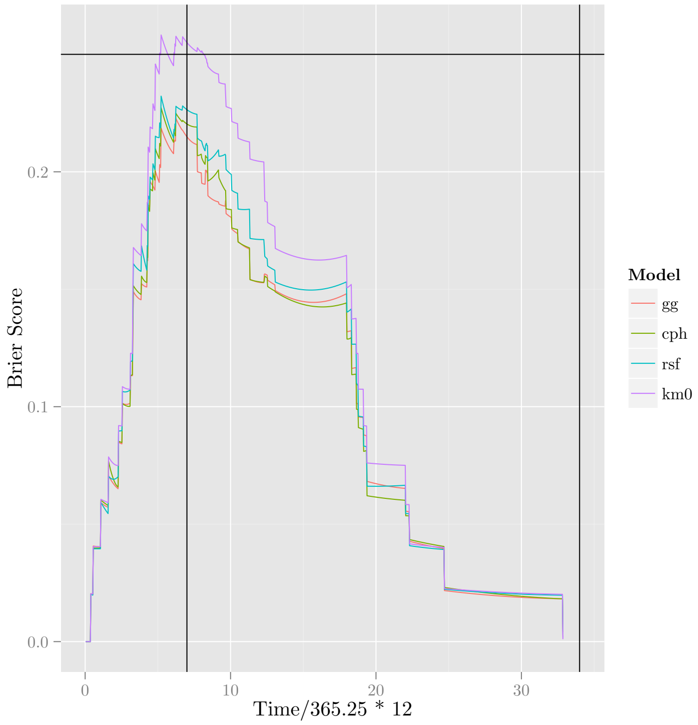


Figure 1.6: Brier score paths for candidate models on the holdout NSWPCN model test set. All models outperformed the no-information KM0 trace from approximately 100 days to 600 days post-diagnosis, and no strong differences were apparent between candidate models.

Both the incident/dynamic TD-ROC, and the IBS, were used to compare model prognostic performance. TD-ROCs were estimated using R package `risksetROC`, and Brier score paths and IBSs were calculated with custom code, following [**?**].

External validation

MSKCC nomogram calculation

The prognostic nomogram for resected pancreas cancer of [**?**] was digitized and transformed into `R` code that produced 12-, 24-, and 36-month disease-specific survival estimates given patient CPVs (see Appendix **??** on page **??**). Predictions for patients with data missing for some nomogram variables were generated by marginalizing over the missing predictors, using the variable distributions in [**?**].

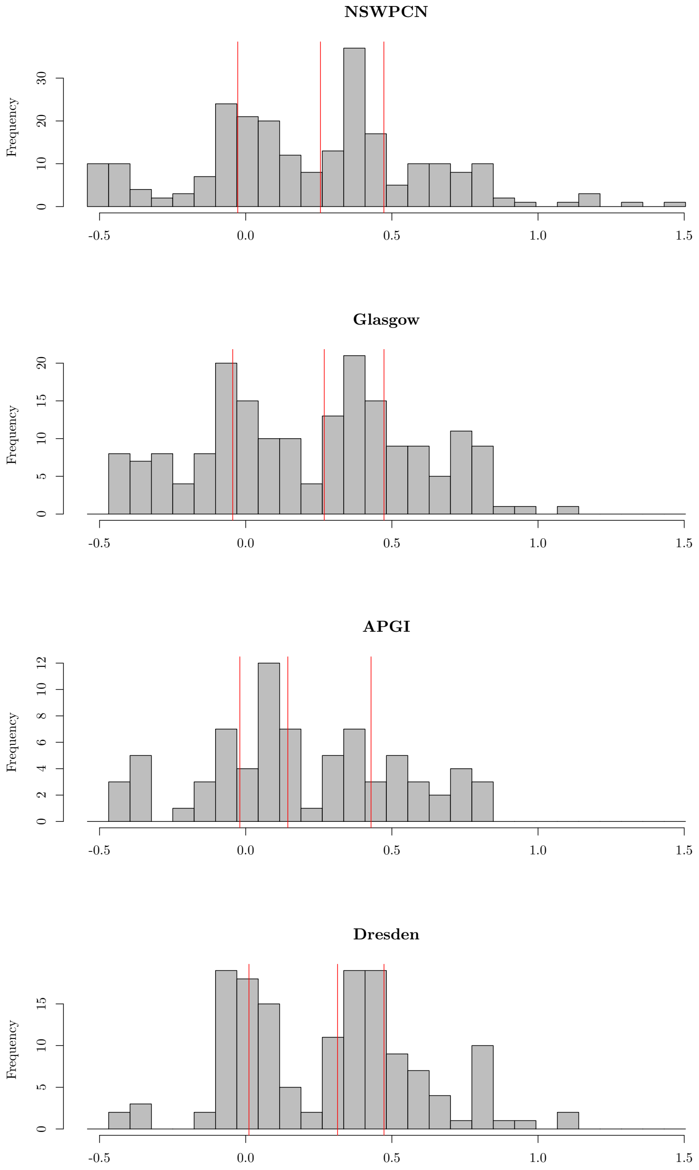


Figure 1.7: Distributions of the PCOP PI in training and validation cohorts. Score distributions were broadly similar in all cohorts, with a generally bimodal form. Empirical quartiles are indicated by red lines.

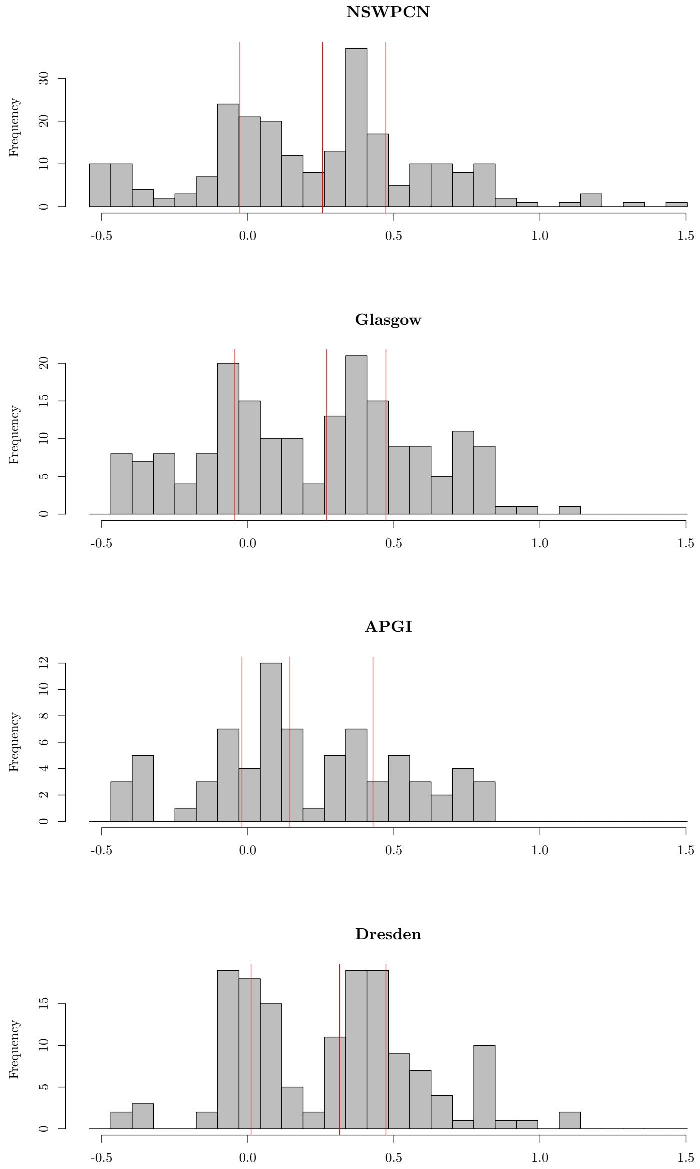
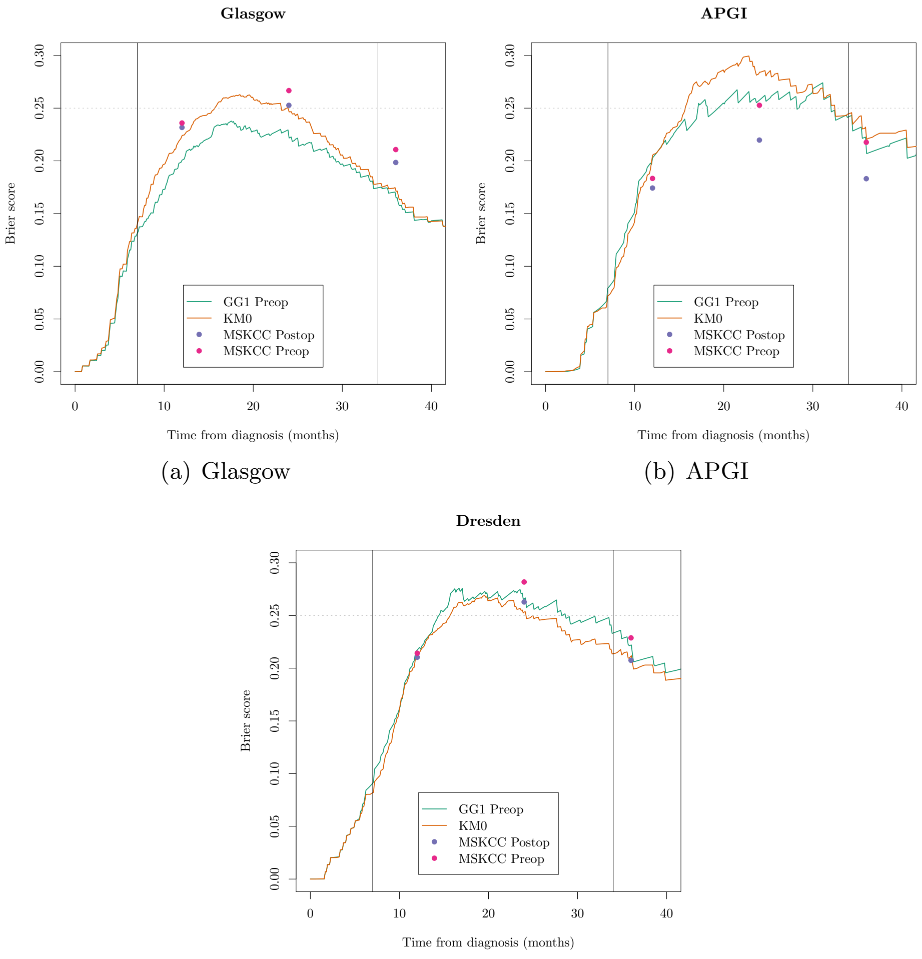


Figure 1.8: Distributions of the PCOP PI in training and validation cohorts. Score distributions were broadly similar in all cohorts, with a generally bimodal form. Empirical quartiles are indicated by red lines.



(a) Glasgow

(b) APGI

(c) Dresden

Figure 1.9: Brier score paths for PCOP on validation cohorts. The Brier score measures overall prognostic error, as a combination of calibration and discrimination; lower values are better, and the worst-case theoretical value is 0.25. Brier score paths over time are shown for outcome predictions by PCOP (green line), and the MSKCC nomogram on preoperative data (pink dots). Also shown is a marginal KM prediction of outcome (orange), and the theoretical no-information Brier score limit (horizontal dotted line) – outcome predictors must be substantially better than both of these to be usefully prognostic. The $7 - 34$ month period in which most patients die is delimited by vertical lines. PCOP is substantially better than either the KM or MSKCC predictors in the Glasgow cohort, but all predictors are equally poor in the APGI and Dresden cohorts.

Table 1.3: Coefficients of a final GG1 fit to the NSWPCN training data, which defines the PCOP. Coefficient estimates are for a generalized gamma survival model [?].

Term		Estimate	SE
β			
(Intercept)		6.7446	0.1489
Sex	= Male	0.3732	0.1508
Tumour location	= Body	−0.2150	0.1223
Size of longest axis	(cm)	−0.0887	0.0302
S100A2	= Positive	−0.3729	0.1235
S100A4	= Positive	−0.3843	0.1045
σ			
(Intercept)		0.7503	0.0493
Sex	= Male	−0.2452	0.1066
λ			
(Intercept)		0.0288	0.2719
Sex	= Male	0.7630	0.3533

(a) Glasgow (b) APGI (c) Dresden

Figure 1.10:

PCOP web application

The `R` `shiny` infrastructure was used to create a simple web application to predict patient outcome using the final PCOP model.