Unified baseline and longitudinal mortality prediction in idiopathic pulmonary fibrosis

Brett Ley¹, Williamson Z. Bradford², Derek Weycker³, Eric Vittinghoff⁴, Roland M. du Bois⁵ and Harold R. Collard¹

Affiliations: ¹Dept of Medicine, University of California San Francisco, San Francisco, CA, USA. ²InterMune Inc., Brisbane, CA, USA. ³Policy Analysis Inc. (PAI), Brookline, MA, USA. ⁴Dept of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA. ⁵Imperial College, London, UK.

Correspondence: Brett Ley, 505 Parnassus Avenue, Box 0111, San Francisco, CA 94143, USA.
E-mail: brett.ley@ucsf.edu

ABSTRACT The Gender-Age-Physiology (GAP) model is a validated, baseline-risk prediction model for mortality in idiopathic pulmonary fibrosis. Longitudinal variables have been shown to contribute to risk prediction in idiopathic pulmonary fibrosis and may improve the predictive performance of the baseline GAP model. Our aims were to further validate the GAP model and evaluate whether the addition of longitudinal variables improves its predictive performance.

The study population was derived from a large clinical trials cohort of patients with idiopathic pulmonary fibrosis (n=1109). Model performance was determined by improvement in the C-statistic, net reclassification improvement, clinical net reclassification improvement, and a goodness-of-fit test.

The GAP model had good discriminative performance with a C-statistic of 0.757 (95% CI 0.750–0.764). However, the original GAP model tended to overestimate risk in this cohort. A novel, easy to use model, consisting of the original GAP predictors plus history of respiratory hospitalisation and 24-week change in forced vital capacity (the longitudinal GAP model) improved model performance with a C-statistic of 0.785 (95% CI 0.780–0.790), net reclassification improvement of 8.5%, clinical net reclassification improvement of 25%, and a goodness-of-fit test of 0.929.

The Longitudinal GAP model, along with the original GAP model, may unify baseline and longitudinal mortality risk prediction in idiopathic pulmonary fibrosis.

@ERSpublications

GAP and longitudinal GAP models may provide simple unified baseline and longitudinal mortality risk prediction in IPF <http://ow.ly/FhYwZ>

For editorial comment see *Eur Respir J* 2015; 45: 1208–1210 [10.1183/09031936.00043915].

This article has supplementary material available from erj.ersjournals.com

Received: Aug 09 2014 | Accepted after revision: Nov 25 2014 | First published online: Jan 22 2015

Support statement: This study was funded by the University of California, San Francisco Nina Ireland Program for Lung Health and InterMune, Inc.

Conflict of interest: Disclosures can be found alongside the online version of this article at erj.ersjournals.com

Copyright ©ERS 2015

Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic progressive fibrotic lung disease with an overall poor prognosis [1]. However, there is substantial heterogeneity in risk of death in individual IPF patients at the time of diagnosis, ranging from <1 year to >10 years [2]. Accurate risk prediction in IPF is important for guiding clinical care.

A simple to use, baseline, clinical-risk prediction model for IPF called the GAP (Gender-Age-Physiology) model has been developed and validated in cohorts of patients seen at tertiary referral centres [3]. The GAP model includes variables (gender, age, physiology, forced vital capacity (FVC) and diffusing capacity of the lung for carbon monoxide (*DLCO*)), which are available to most clinicians at the time of initial evaluation. However, there are likely aspects of the disease, not captured by the GAP model, which are informative regarding risk prediction, in particular longitudinal disease behaviour. A second risk prediction model, a model by DU BOIS *et al.* [4] (the du Bois *et al.* model), which was developed in a large clinical trial cohort, incorporated two longitudinal variables, these being change in FVC over a 24-week period and a recent history of respiratory hospitalisation. Recently this latter model has been updated to include 6-min walk test (6MWT) variables [5]. Importantly, the du Bois *et al.* model may not be suitable for risk assessment in severe patients or patients who do not have historical data available. For both the GAP and the du Bois *et al.* models, it is unknown how well they perform in patient populations distinct from those in which they were derived.

The aims of this study were to assess the robustness of the GAP baseline model across diverse patient populations, in particular clinical trial cohorts, and to evaluate the predictive value of GAP-based models incorporating additional baseline and longitudinal variables (a combined GAP/du Bois *et al.* model), thereby simplifying baseline and longitudinal mortality risk prediction in IPF.

Methods

Study population

The source population included all randomised subjects from a clinical trial of interferon γ 1b in IPF (GIPF-007, $n=826$), and all randomised placebo subjects from two clinical trials of pirfenidone in IPF (PIPF-004 and PIPF-006, $n=347$). These patients and protocols have been well described previously [6, 7]. All eligible subjects had IPF by consensus criteria [8]. All subjects enrolled in these three trials had mild-to-moderate physiological impairment as defined by an FVC ≥ 50 –55% pred and a *DLCO* $\geq 35\%$ pred.

The current study cohort included all subjects in the source population who had a 24-week trial visit (fig. 1a). This was done to provide historical data for the longitudinal predictor variables (e.g. 24-week change in FVC) and to provide a more balanced distribution of disease severity.

Predictor variables

Predictor variables were pre-specified based on clinical relevance and evidence supporting their association with prognosis in IPF. These included gender, age, baseline FVC, baseline *DLCO* (and an indicator for the inability to perform the *DLCO* test at baseline), baseline 6-min walking distance (6MWD) in metres, baseline dyspnea score on the University of California San Diego Shortness of Breath Questionnaire (UCSD SOBQ), lowest oxygen saturation level on the 6MWT (6MWT desaturation) in per cent saturation,

FIGURE 1 Study design including a) the study population and b) the study period. GIPF-007: clinical trial of interferon γ 1b in idiopathic pulmonary fibrosis (IPF); PIPF-004/-006: clinical trials of pirfenidone in IPF.

use of long-term oxygen therapy (LTOT), occurrence of respiratory hospitalisation in the last 24 weeks, 24-week relative change in FVC, 24-week relative change in *DLCO*, 24-week absolute change in 6MWD in metres, and 24-week absolute change in dyspnoea score. Other, baseline characteristics that were collected, but not evaluated as predictor variables, included ethnicity (white, yes/no), body mass index, smoking status (ever/never) and history of surgical lung biopsy (yes/no).

Primary outcome

The primary outcome was time from baseline (defined as the date of subjects' week 24 visit in the parent clinical trial) to death from any cause (fig. 1b). Vital status was available on all subjects for the duration of the parent clinical trials. Subjects were right censored at the time of loss to follow-up, at the end of follow-up, and at the time of lung transplantation.

Statistical analysis

Analyses were done using STATA 13 (StataCorp, College Station, TX, USA). Prediction models were based on Cox proportional hazards models. First, we evaluated the predictive performance of the original, fully specified, GAP model in the study cohort, including the GAP calculator, which is the GAP model that uses variables in continuous form, and the GAP index, which is the GAP model that uses variables in a categorised form with point-score assignments [3]. Model performance was evaluated by its components, discrimination and calibration. Discrimination is the ability of a model to discriminate those with an outcome from those without an outcome. For survival data (*i.e.* time-to-event data) discrimination is measured using the C-statistic, which ranges from 0 to 1.0, with 0.5 indicating no predictive discrimination and 1.0 indicating perfect discrimination [9]. The C-statistic is analogous to the area under the receiver operator curve used for binary outcomes. In general, a C-statistic value of 0.70–0.80 is considered good, 0.80–0.90 is excellent, and >0.90 is outstanding. Calibration is the determination of how closely model-predicted outcomes approximate actual outcomes. We evaluated calibration by comparing model-predicted mortality to observed mortality; observed mortality was estimated using the Kaplan–Meier method. The GAP and du Bois *et al.* models were “re-fit” by generating new model coefficients to optimise performance in the study cohort. For both the “refit” models, discriminative performance was re-evaluated. To evaluate global model fit, we applied a pseudo-Hosmer–Lemeshow goodness-of-fit (GOF) test, where a higher value indicates a better model fit [10]. Discrimination is considered the primary measure of model predictive performance because, in contrast to calibration and global model fit, it cannot be improved with adjustment or “re-calibration” [9]. Therefore, the C-statistic was our primary measure for comparing models.

Next, a screening procedure was performed to rank all potential models that include the individual GAP predictors, plus one to four additional variables (“base model” approach) by the cross-validated C-statistic [9]. Novel (*i.e.*, non-original GAP) predictors were modelled in continuous/linear and categorical/dichotomous forms. Cut-off values for categorised/dichotomised predictors were based on prior studies and/or clinical convention [11–15]. Models with GAP plus novel predictors were compared to the GAP model by the change in the C-statistic, net reclassification improvement (NRI) [16], and clinical NRI (cNRI) [17, 18]. Risk strata for NRI were based on three strata of the 1-year mortality risk: <10% (low risk), 10–20% (intermediate risk), and >20% (high risk). These strata were chosen in light of the average 1-year mortality risk after lung transplantation of 20% [19]. NRI estimates the net proportion of surviving or dying patients overall correctly reclassified into a lower or higher risk stratum, respectively, by the new model. cNRI focuses on the net proportion of patients at intermediate risk (presumably a less clinically-informative stratum) correctly reclassified into the low or high risk strata. After a few “good” variables are included in a model, the C-statistic is relatively insensitive in detecting the added value afforded by additional predictor variables; meaning, the C-statistic may change very little, yet important (*i.e.* clinically useful) improvements may be afforded by the additional predictor variable. In this instance, reclassification statistics are often employed to demonstrate this benefit [20]. Bootstrap re-sampling with 500 repetitions was used to calculate 95% bias-corrected confidence intervals. Finally, a novel “longitudinal GAP” model was selected to balance optimum discrimination with practical considerations, such as ease of use in the clinical setting. This model was then extended to a point-score model using methods previously described [4].

Results

Cohort characteristics and validation of the GAP model

There were 1109 subjects included in the study cohort. Table 1 shows the characteristics of the baseline cohort. Median (range) follow-up was 1.1 years (0.01–2.36 years), during which 128 deaths and 25 lung transplantations occurred. In this cohort, the original GAP calculator had a C-statistic of 0.698 (95% CI 0.647–0.749) and the original GAP index had a C-statistic of 0.676 (95% CI 0.625–0.728), similar to the

TABLE 1 Cohort characteristics at baseline in the preceding 24 weeks

Characteristic	Value
Patients	1109
Age	66.2±7.7
Male	791 (71.3)
Ethnicity white	1050 (94.7)
Body mass index Kg·m⁻²	29.9±4.7
Smoking status ever	769 (69.3)
Oxygen use	192 (17.3)
Surgical lung biopsy	602 (54.3)
FVC % predicted	71.1±15.2
24-week relative decline FVC ≥10%	213 (19.2)
D_LCO % predicted	44.4±11.4
Missing	25 (2.3)
6-min walking distance m	376.5±123.4
UCSD SOBQ	37.2±23.5
Respiratory hospitalisation in the preceding 24 weeks	69 (6.2)
GAP stage	
1	443 (39.9)
2	558 (50.3)
3	108 (9.7)

Data are presented as n, mean±SD, or n (%). FVC: forced vital capacity; D_LCO: diffusing capacity of the lung for carbon monoxide; 6 UCSD SOBQ: University of California San Diego Shortness of Breath Questionnaire; GAP: Gender-Age-Physiology.

discriminative performance found in the previous clinical cohorts (C-statistic 0.695 and 0.697, respectively) [3]. The original GAP models consistently over-estimated mortality risk in this clinical trial cohort, especially in low to moderate risk groups, indicating poor calibration (table 2).

After re-fitting the GAP model to the study cohort, the C-statistic increased to 0.757 (95% CI 0.750–0.764); GOF was 0.539. Of the individual GAP predictors, age and gender were not significantly associated with mortality in the current study cohort; FVC, D_LCO, and the inability to perform the D_LCO test were the most strongly associated variables (table S1). The “refit” du Bois *et al.* model demonstrated similar performance to the “refit” GAP model (C-statistic 0.747, 95% CI 0.740–0.755; GOF 0.743).

Association of novel variables with mortality

On unadjusted (bivariate) analysis of the novel predictor variables significant associations with mortality were found for: respiratory hospitalisation in the prior 24 weeks; dyspnoea severity, as measured by the baseline UCSD SOBQ and 24-week change in UCSD SOBQ (modelled continuously); baseline 6MWD and 24-week change in 6MWD (modelled continuously); 6MWT desaturation; 24-week change in FVC (modelled continuously and dichotomously at >10%); and 24-week change in D_LCO (modelled continuously) (table S2). The use of long-term oxygen therapy was not significantly associated with mortality. After adjustment for the GAP variables, significant associations remained for: respiratory hospitalisation (p<0.001); baseline UCSD SOBQ (p=0.014); baseline 6MWD (p<0.001) and 24-week change in 6MWD (p<0.001); 6MWT desaturation (p=0.001); and 24-week change in FVC (modelled continuously, p=0.001) (fig. 2 and table S2). Variables that were no longer independently associated with mortality after adjustment for the GAP model included: 24-week change in FVC (modelled dichotomously at ≤–10%, p=0.077), 24-week change in UCSD SOBQ (p=0.248), and a 24-week change in D_LCO (p=0.263).

Additional predictive value of novel variables

Individual variables that significantly improved discriminative performance when added to the GAP model were respiratory hospitalisation and UCSD SOBQ, while those that non-significantly improved performance were 6MWD and 24-week change in FVC (continuous variable) (table 3). Respiratory hospitalisation (and 6MWD) also demonstrated significant improvements in the NRI and cNRI. The GAP plus respiratory hospitalisation model demonstrated the best global model fit (GOF 0.998). Predictors that did not improve discriminative performance of the GAP model were a 24-week decline in FVC >10%, 24-week change in SOBQ, use of LTOT, 24-week change in 6MWD, 24-week change in D_LCO, and a 6MWT desaturation.

TABLE 2 Calibration of the original Gender-Age-Physiology (GAP) models in the clinical trial cohort

GAP	Patients n	GAP calculator			GAP index		
		Predicted mortality [#]	Observed mortality [¶]	Ratio ⁺	Predicted mortality [#]	Observed mortality [¶]	Ratio ⁺
1-year[§]							
Stage 1	444	8.5	4.6	0.54	5.6	4.6	0.82
Stage 2	561	17.6	8.4	0.48	16.2	8.4	0.52
Stage 3	108	29.1	30.6	0.95	39.2	30.6	0.78
2-year[§]							
Stage 1	444	17.4	11.3	0.65	10.9	11.3	1.04
Stage 2	561	33.9	22.1	0.65	29.9	22.1	0.74
Stage 3	108	52.0	45.9	0.88	62.1	45.9	0.74

Data are presented as % unless otherwise specified. [#]: based on predictions from models reported in LEY *et al.* [3]; [¶]: based on Kaplan–Meier estimates for all-cause mortality in the study cohort; ⁺: ratio of observed to predicted mortality; [§]: stages as defined by LEY *et al.* [3].

Table 4 summarises the discriminative performance of models constructed from the GAP model plus multiple additional variables. The addition of respiratory hospitalisation, 24-week change in FVC, and baseline UCSD SOBQ resulted in the largest improvement in discrimination (C-statistic 0.788, 95% CI 0.783–0.794; change in C-statistic 0.032, 95% CI 0.023–0.040). A simpler model consisting of the GAP model plus respiratory hospitalisation and 24-week change in FVC had comparable performance (C-statistic 0.785, 95% CI 0.780–0.790). Both models demonstrated significant risk reclassification improvement (NRI and cNRI) and excellent global fit. The simpler model was further specified as an easy-to-use point-score model (tables S3–S5).

Discussion

This study validates the GAP model in the clinical trial population and unifies the previously published GAP and du Bois *et al.* risk prediction models into a single risk prediction approach that includes a baseline model (the GAP model) and a longitudinal model (the Longitudinal GAP model) (fig. 3). We hope that this provides clinicians and clinical trialists with a simple and easy to use risk-prediction approach, which can be applied to both baseline and longitudinal risk assessment.

FIGURE 2 Forest plot demonstrating the association of individual variables with mortality after adjustment for the Gender-Age-Physiology (GAP) model. The lines represent 95% confidence intervals. UCSD SOBQ: University of California San Diego Shortness of Breath Questionnaire; 6MWD: 6-min walking distance; 6MWT: 6-min walk test; FVC: forced vital capacity; DLco: diffusing capacity of the lung for carbon monoxide; HR: hazard ratio. [#]: preceding 24 weeks; [¶]: per 10 units; ⁺: per 50 m; [§]: per 5% decrease; ^f: continuous, per 5%; ^{##}: binary >10%.

TABLE 3 Additional predictive values of individual variables compared to the Gender-Age-Physiology (GAP) model

Model	C-statistic (95% CI)	Change in C-statistic compared to GAP (95% CI)	NRI (95% CI)	cNRI (95% CI)	GOF
GAP+RH	0.776 (0.771–0.780)	0.019 (0.011–0.027)	7.7 (1.3–19.2)	23.3 (7.1–60.5)	0.998
GAP+UCSD SOBQ	0.766 (0.761–0.772)	0.009 (0.001–0.018)	4.7 [–4.9–12.1]	7.1 [–11.8–18.5]	0.797
GAP+6MWD	0.765 (0.753–0.776)	0.008 [–0.005–0.021]	11.4 (3.7–25.1)	30.8 (8.7–56.3)	0.025
GAP+FVC 24-week change	0.762 (0.753–0.776)	0.005 [–0.004–0.021]	3.8 [–4.8–14.5]	11.8 [–5.8–38.7]	0.792
GAP alone	0.757 (0.750–0.764)				0.539
GAP+FVC decline >10%	0.756 (0.748–0.764)	–0.001 [–0.011–0.010]	3.5 (0.4–14.3)	5.1 [–1.3–21.4]	0.802
GAP+UCSD SOBQ 24-week change	0.755 (0.748–0.761)	–0.002 [–0.012–0.007]	8.3 (1.0–19.0)	5.6 [–6.7–19.4]	0.904
GAP+LTOT	0.753 (0.745–0.761)	–0.004 [–0.014–0.007]	–0.2 [–6.3–4.0]	–1.4 [–15.3–2.5]	0.278
du Bois <i>et al.</i>[#]	0.747 (0.740–0.755)				0.743
GAP+6MWD 24-week change	0.744 (0.737–0.751)	–0.013 [–0.022–0.003]	6.5 [–2.0–17.1]	16.0 [–10.6–36.0]	0.041
GAP+DLco 24-week change	0.739 (0.731–0.747)	–0.018 [–0.028–0.007]	7.3 (1.0–17.5)	1.5 [–16.5–9.3]	0.885
GAP+6MWT desaturation	0.732 (0.703–0.762)	–0.025 [–0.055–0.005]	6.2 (0.6–17.1)	17.8 (5.9–37.9)	0.175

NRI: net reclassification improvement; cNRI: clinical NRI improvement; GOF: goodness-of-fit test; RH: respiratory hospitalisation; UCSD SOBQ: University of California San Diego Shortness of Breath Questionnaire; 6MWD: 6-min walking distance; FVC: forced vital capacity; LTOT: long-term oxygen therapy; DLco: diffusing capacity of the lung for carbon monoxide; 6MWT: 6-min walk test; [#]: du Bois *et al.* model involves age, 24-week change in FVC, baseline FVC and history of RH in the preceding 24 weeks.

Our results provide some interesting insights into the differences in clinical trial and clinical practice-based cohorts. First, the individual predictors in the GAP model contribute differently to risk, depending on the cohort. In the current clinical trial cohort, demographic variables (age and gender), contribute less, with risk determined almost entirely from FVC and DLCO. Second, a comparison of model-predicted and observed mortality risk shows that GAP mortality risk estimates, derived from the clinical care setting, substantially overestimate mortality risk observed in the clinical trial setting. We suspect that these differences may be due to unmeasured confounders present in clinical cohorts, but not clinical trial cohorts. One likely confounder is comorbidity (e.g. cardiovascular disease and cancer), which may associate with age and gender and increase overall mortality. In other words, patients with clinically significant comorbidities may be less common in clinical trial cohorts compared to clinical cohorts (as a consequence of direct and indirect exclusion), making age and gender (markers of this comorbidity) less predictive and death less frequent. Regardless of the reason for these findings, they suggest that risk prediction models may need to be estimated and calibrated separately for clinical and clinical trial cohorts.

In evaluating additional predictors, the variable “respiratory hospitalisation in the preceding 24 weeks” appears to provide the greatest improvement in performance over the original GAP model, increasing the risk of subsequent death by more than three-fold (HR 3.3), independent of GAP-adjusted risk. Its addition provides significant improvement in discrimination (C-statistic increases nearly 2%) and risk reclassification.

TABLE 4 Additional predictive values of multiple variables compared to the Gender-Age-Physiology (GAP) model

Model	C-statistic (95% CI)	Change in C-statistic compared to GAP (95% CI)	NRI (95% CI)	cNRI (95% CI)	GOF
GAP+RH+FVC 24-week change+UCSD SOBQ	0.788 (0.783–0.794)	0.032 (0.023–0.040)	17.6 (9.1–30.7)	31.7 (12.2–57.1)	0.715
GAP+RH+FVC 24-week change+UCSD SOBQ+6MWD	0.788 (0.776–0.800)	0.031 (0.018–0.045)	15.1 (1.0–23.1)	35.7 (10.0–51.4)	0.506
GAP+RH+FVC 24-week change	0.785 (0.780–0.790)	0.028 (0.020–0.036)	8.5 [–1.0–18.8]	25.0 (4.9–50.3)	0.929
GAP+RH+UCSD SOBQ	0.783 (0.778–0.788)	0.026 (0.018–0.035)	16.1 (9.1–29.6)	29.0 (10.1–58.2)	0.979
GAP+RH+6MWD	0.780 (0.770–0.790)	0.023 (0.011–0.035)	12.6 (4.0–24.7)	35.4 (13.0–59.5)	0.016
GAP+FVC 24-week change+6MWD	0.767 (0.755–0.778)	0.010 [–0.003–0.023]	13.1 (2.4–27.9)	32.6 (4.7–58.3)	0.664
GAP alone	0.757 (0.750–0.764)				0.539
du Bois <i>et al.</i>[#]	0.747 (0.740–0.755)				0.743
GAP+UCSD SOBQ+6MWD	0.746 (0.714–0.777)	–0.011 [–0.043–0.021]	11.9 (3.9–24.1)	32.2 (3.3–61.4)	0.083

NRI: net reclassification improvement; cNRI: clinical NRI improvement; GOF: goodness-of-fit test; RH: respiratory hospitalisation; FVC: forced vital capacity; UCSD SOBQ: University of California San Diego Shortness of Breath Questionnaire; 6MWD: 6-min walking distance; [#]: du Bois *et al.* model included: age, 24-week change in FVC, baseline FVC, and history of RH in the preceding 24-weeks.

FIGURE 3 Proposed clinical application of Gender-Age-Physiology (GAP) models for baseline and longitudinal mortality risk prediction in idiopathic pulmonary fibrosis. PFTs: pulmonary function tests; FVC: forced vital capacity; DLCO: diffusing capacity of the lung for carbon monoxide; Δ FVC: 24-week relative change in FVC. #: history of hospitalisation for respiratory worsening in the previous 24 weeks.

More than 23% of patients assigned to the “intermediate” risk group by the original GAP model were correctly “re-assigned” to higher or lower risk groups by the inclusion of respiratory hospitalisation. A history of respiratory hospitalisation has clear, face validity and precedent as a predictor of poor outcome in IPF [9, 10], and it is not difficult for clinicians to ascertain in real-world scenarios.

The variable “relative change in FVC over 24 weeks” (modelled as a continuous variable) also improves discriminative performance when added to the GAP model, although less so than respiratory hospitalisation. Decline in FVC is a well described prognostic factor in IPF [11–14], and is a common means of defining disease progression. The major limitation of change in FVC, as a practical predictor variable, is that it requires patients come to a clinic for repeated testing. However, home monitoring of FVC through the use of hand-held spirometers may make this less of an issue.

A central challenge in developing useful clinical prediction models is balancing optimal statistical performance with real-world practicality. Although not the top performing model (table 4), we believe that the Longitudinal GAP model provides the best balance of improvement in discriminative performance (increase in C-statistic of ~3%) and risk reclassification (cNRI of 25%) with comprehensive, multidimensional risk assessment, feasibility, and ease of use. The top performing model included the UCSD SOBQ, which we felt added very little to discriminative performance (additional 0.4% increase) and added substantial complexity to the model, as the UCSD SOBQ is not uniformly administered in clinical and research settings.

As previously stated, the Longitudinal GAP model also nicely merges the GAP [3] and du Bois *et al.* models, improving upon their individual predictive performance, and providing a unified baseline and longitudinal mortality risk prediction system in IPF. Of note, the du Bois *et al.* model was recently updated to include 6MWD and 24-week change in 6MWD with a C-statistic similar to the Longitudinal GAP model (C-statistic 0.80) [5]. As for the UCSD SOBQ, we felt that 6MWT parameters added complexity to the clinical evaluation and, in this study, did not appreciably improve model performance in the context of the GAP model.

The major strengths of this study are its use of high-quality, prospectively collected data and its sophisticated risk prediction modelling methodology. With respect to the latter, we believe strongly that the evaluation of hazard ratios and p-values alone (a common approach used in the literature) is not sufficiently robust to identify, and appropriately, determine the additional predictive value of novel variables. For example, in our data oxygen desaturation on the 6MWT maintains statistical independence after adjustment for the GAP model ($p=0.001$), but its addition to the GAP model actually worsens discriminative performance (decrease in C-statistic of 2.5%). Future risk prediction models should use less biased methodology, where component variables are chosen based on overall model performance rather than individual associations with risk.

In this study, we did not evaluate the performance of other novel combinations of predictors outside the GAP construct. It is possible that models replacing one or more of the GAP predictors could have better performance. We believe using the GAP model as a “base model” on which to build provides practical value (simplicity and flexibility) and so chose to require the inclusion of these variables. In addition, even though models developed in this cohort build from a well-validated model, the Longitudinal GAP model still requires external validation. We were unable to evaluate the performance of non-clinical predictors of interest, such as blood biomarkers. Lastly, due to the limited follow-up time afforded by a clinical trial cohort, the Longitudinal GAP model does not predict outcomes beyond 2 years.

In conclusion, we suggest that a unified approach to mortality risk prediction modelling in IPF, based on the GAP and Longitudinal GAP models, provides a simple and easy to use tool for clinicians and clinical

trialists. We anticipate that clinicians may eventually adopt a two-tiered approach to risk prediction in IPF patients, with an initial assessment of mortality risk using the GAP model and subsequent assessment using the Longitudinal GAP model. Clinical trialists may use the Longitudinal GAP model to further refine cohort enrichment for patients at increased risk for death. However, before widespread adoption, the Longitudinal GAP model should be further validated and calibrated separately for clinical and experimental trial cohorts. Additionally, future research should evaluate the value of the GAP and Longitudinal GAP models in predicting risk of non-mortality outcomes. If these models also inform the risk of disease progression or acute exacerbation, their use in clinical practice and clinical trial development will be greatly enhanced.

Acknowledgements

We would like to acknowledge Peter Gaccione and Mark Atwood from Policy Analysis Inc. Brookline, MA, USA, for their assistance with data processing and analysis.

References

- 1 Raghu G, Collard HR, Egan JJ, *et al.* An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011; 183: 788–824.
- 2 Nathan SD, Shlobin OA, Weir N, *et al.* Long-term course and prognosis of idiopathic pulmonary fibrosis in the new millennium. *Chest* 2011; 140: 221–229.
- 3 Ley B, Ryerson CJ, Vittinghoff E, *et al.* A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med* 2012; 156: 684–691.
- 4 du Bois RM, Weycker D, Albera C, *et al.* Ascertainment of individual risk of mortality for patients with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011; 184: 459–466.
- 5 du Bois RM, Albera C, Bradford WZ, *et al.* 6-minute walk distance is an independent predictor of mortality in patients with idiopathic pulmonary fibrosis. *Eur Respir J* 2014; 43: 1421–1429.
- 6 King TE, Albera C, Bradford WZ, *et al.* Effect of interferon gamma-1b on survival in patients with idiopathic pulmonary fibrosis (INSPIRE): a multicentre, randomised, placebo-controlled trial. *Lancet* 2009; 374: 222–228.
- 7 Noble PW, Albera C, Bradford WZ, *et al.* Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials. *Lancet* 2011; 377: 1760–1769.
- 8 American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS). *Am J Respir Crit Care Med* 2000; 161: 646–664.
- 9 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361–387.
- 10 Parzen M, Lipsitz SR. A global goodness-of-fit statistic for Cox regression models. *Biometrics* 1999; 55: 580–584.
- 11 du Bois RM, Weycker D, Albera C, *et al.* Six-minute-walk test in idiopathic pulmonary fibrosis: test validation and minimal clinically important difference. *Am J Respir Crit Care Med* 2011; 183: 1231–1237.
- 12 Zappala CJ, Latsi PI, Nicholson AG, *et al.* Marginal decline in forced vital capacity is associated with a poor outcome in idiopathic pulmonary fibrosis. *Eur Respir J* 2010; 35: 830–836.
- 13 du Bois RM, Weycker D, Albera C, *et al.* Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference. *Am J Respir Crit Care Med* 2011; 184: 1382–1389.
- 14 Kupferberg DH, Kaplan RM, Slymen DJ, *et al.* Minimal clinically important difference for the UCSD Shortness of Breath Questionnaire. *J Cardiopulm Rehabil* 2005; 25: 370–377.
- 15 Eaton T, Young P, Milne D, *et al.* Six-minute walk, maximal exercise tests: reproducibility in fibrotic interstitial pneumonia. *Am J Respir Crit Care Med* 2005; 171: 1150–1157.
- 16 Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011; 30: 11–21.
- 17 Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biom J* 2011; 53: 237–258.
- 18 Cook NR. Comments on 'evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M. J. Pencina *et al.*, *Statistics in Medicine* (DOI: 10.1002/sim.2929). *Stat Med* 2008; 27: 191–195.
- 19 Thabut G, Christie JD, Ravaud P, *et al.* Survival after bilateral *versus* single-lung transplantation for idiopathic pulmonary fibrosis. *Ann Intern Med* 2009; 151: 767–774.
- 20 Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21: 128–138.