SUPPLEMENTARY APPENDIX

MATCHING PROCESS

Doctors in Commissions' complaint databases were matched to doctors in the AMPCo Direct database in order to obtain a complete set of demographic variables—including including, age, sex, specialty and subspecialty, primary practice location—on all doctors in the study sample. The matching was probabilistic. It was done with FRIL linkage software (version 2.1.4, Emory University) and was based on doctor name, specialty, and practice postcode. The match rate across Commissions averaged 85% (range, 75% to 87%).

For doctors named in complaints who could not be matched to the AMPCo Direct list, we sought to retrieve any missing doctor-level information from publicly-accessible sources, including the national register of medical practitioners(1), newspaper obituaries and other media reports, and the newsletters of colleges, societies and other professional associations.

After both matching to the AMPCo Direct database and the manual addition of missing variables, 97% of doctors in the sample frame entered the study dataset.

DISTRIBUTIONAL/CLUSTERING ANALYSIS

Statistics on complaint clustering among doctors were calculated as follows:

(a) Number of complaints	(b) Number of doctors with complaint count specified in column (a)	(c) % of all complained-against doctors with complaint count (strictly) exceeding that specified in column (a)	(d) % of all practicing doctors with complaint count specified in column (a)	(e) % of all complaints
1 complaint	x	(1 – (x / Q))*100	(1 – (x / R)) *100	[1 – ((1 * x) / P)] * 100
2 or more complaints	y	(1 – (y / Q)) *100	(1 – (y / R)) *100	[1 – (((1*x) + (2 * y)) / P)] *100
n or more complaints	z	(1 – (z / Q)) *100	(1 – (z / R)) *100	[1 – (((1*x) + (2 * y) +…+ (n * z)) / P)] *100

where P is the total complaint count, Q is the total number of complained-against doctors with and R is the total number of practicing doctors.

The statements in the manuscript reporting distribution of complaints among doctors were based on the doctor percentages calculated in column (c) and column (d), respectively, and the corresponding complaints percentages (i.e. same row) in column (e).

All values in the above table came directly from the analytic dataset, except the total number of practicing doctors (R). We obtained R from medical workforce data series[1] that is published annually by the Australian Institute of Health and Welfare, the official keeper of health statistics in Austalia.

Some additional information about the R value used is provided below:

- We elected to make R the total number of doctors in 2006, the mid-point of our study.

- To test how sensitive the distributional statistics were to the choice of the year 2006, we recalculated them using the R values from 2004 and 2008. Year-to-year variations in R had negligible effects on the distributional statistics. For example, on the basis of the 2006 R value, we report that around 1% of doctors accounted for 25% of all complaints.

The precise doctor percentages by year are: 1.0% (using R from 2004); 0.9% (using R from 2006); and 1.1% (using R from 2008). Similarly, we report that about 3% of doctors accounted for 49% of all complaints. The precise doctor percentages by year are: 3.6% (using R from 2004); 3.3% (using R from 2006); and 3.9% (using R from 2008). The increase in 2008 appears to be due to a minor change in the way that the Australian Institute of Health and Welfare tallied doctors, rather than a consequence of a substantial year-to-year change in the total number of doctors.

- To create the appropriate R value for purposes of our study, several adjustments had to be made to the raw doctor totals by state and territory reported in the AIHW labour force report. Specifically:

    - We subtracted the total number of doctors in South Australia; this was the state that did not participate in the study.

    - We subtracted the total number of doctors in New South Wales because complaints from this state were not used in the distributional statistics. The "exposure period" on which the distributional statistics are based is 10 years and, as we explain in the manuscript: *"Data from New South Wales was not included in these plots because the complaints window there spanned only five years."*

    - Although Commissioners' have regulatory authority over both private and public health services in their jurisdictions, there were challenges in some jurisdictions in capturing complaints arising from the public hospital setting. In four of the six jurisdictions whose data was used for the distributional statistics, Commissioners have a practice of opening public hospital complaint files in the name of the hospital, not individual clinicians. This inhibited our ability to identify complaints against doctors in these jurisdictions when they arose in the public hospital setting. The other two jurisdictions routinely open complaint files in the name of any clinician complaint against, regardless of where the care is rendered.

      We accounted for this variation in "exposure" to complaints in constructing R. For the four states in which our complaints data did not include public hospital complaints, the states' contribution to R was based on a count of doctors who were employed (at least some of the time) in private practice (this included all General Practitioners). Doctors in full-time public practice were excluded. For the two states in which our complaints data included public hospital complaints, we used counts of all doctors.

- Finally, we note that our method of calculating R by summing doctor totals in each state and territory may have resulted in a slight over-count, because some doctors practice in multiple jurisdictions. In the absence of a national registration scheme (which did not commence in Australia until July 2010), we were unable to quantify the extent of this over count. Its effect would be to render our estimates of complaint clustering among doctors a lower bound on the true extent of clustering.

Multivariable Analysis

Choice of statistical model

For our main analysis, we used a model that had a common baseline hazard and defined time as time since entry into the study (from the date of a doctor's first complaint). This model is generally referred to as the Anderson-Gill (AG) model.(2) In addition to assuming a common baseline hazard, this model treats all failures (initial and subsequent) as exchangeable and independent, conditional on the covariates. The general form of the hazard function for the AG model is

$$h_j(t|x_j) = h_0(t)exp(x_j\beta)$$

where $h_0(t)$ is the baseline hazard and $x_j\beta = \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n$ estimates how the hazard changes as a function of specified covariates.

One criticism of the AG model is that its assumption about a common baseline hazard may not be appropriate in all situations. An alternative approach would be to use a model that allowed the baseline hazard to vary with the occurrence of successive "events". (In our analysis, "events" refer to the number of complaints a doctor experienced.) This type of model is commonly referred to as a Prentice, Williams and Peterson counting process model(3) (PWP-CP); it is a conditional model in the sense that the subject can only be at risk of the *k*th event if the *k* – 1 event has occurred.(4) The hazard function for this model is

$$h_{jk}(t|x_{jk}) = h_{0k}(t)exp(x_{jk}\beta), \quad k = 1, 2, \dots$$

where $h_{0k}(t)$ is the stratum-specific baseline hazard.

An important disadvantage of the PWP-CP model in the context of our study is that it precludes direct estimation of hazard ratios showing the effect of number-of-previous-complaints. This is problematic because we hypothesized that this previous complaints variable would be a key predictor of doctors' risk of subsequent complaints, and therefore sought to quantify the magnitude of its effect. Nonetheless, in sensitivity analyses described below, we fit a PWP-CP model to our data and compare estimates from it to the AG model used in the main analysis. (This particular sensitivity analysis is not reported in the manuscript due to space constraints and its technical nature for a general medical audience).

Estimation method

AG models can be estimated with Cox regression or with fully parametric models. We used a fully parametric approach for two main reasons: (1) it allows estimation of the shape of the baseline hazard; and (2) it gives smooth estimates of the baseline hazard and survivor functions. Neither of these is possible with Cox models.

Distribution of time

Estimates from parametric models can be sensitive to choices made about the type of distribution imposed in time in the model. The standard options are the Gompertz distribution and the Weibull distribution. When the distribution of time is assumed to follow a Gompertz distribution, the hazard function for the AG model is

$$h_j(t|x_j) = exp(\gamma t)exp(x_j\beta)$$

where $exp(\gamma t)$ is the baseline hazard function and γ is an ancillary parameter estimated from the data. When the distribution of time is assumed to follow the Weibull distribution, the hazard function for the AG model is

$$h(t|x_j) = pt^{p-1}exp(x_j\beta)$$

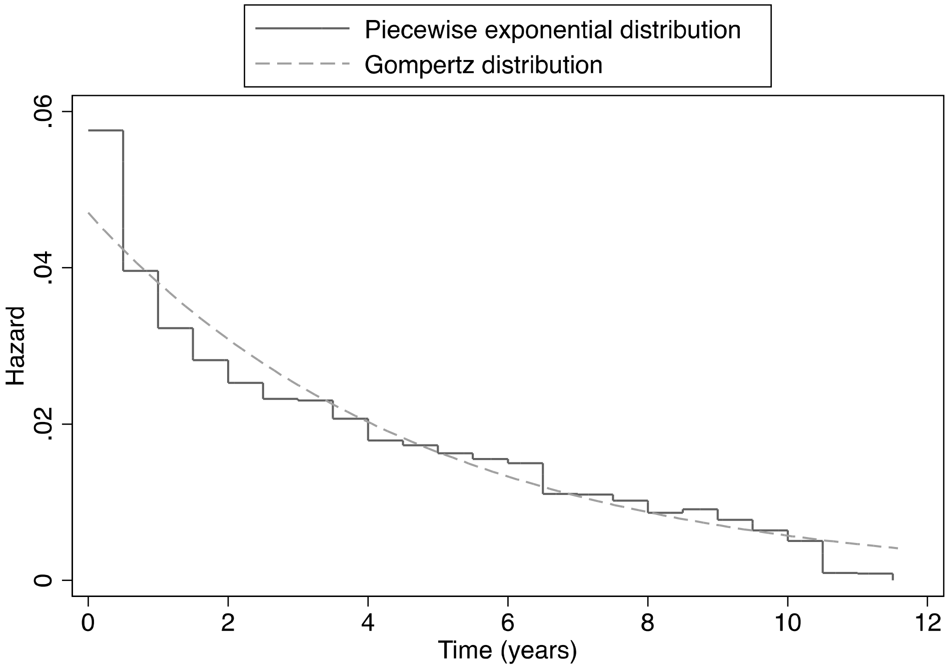
where pt^{p-1} is the baseline hazard function and p is the shape parameter, which is also estimated from the data.

To determine the most appropriate distribution for the baseline hazard, we fit two AG models to the data; one model used a Gompertz distribution, the other used a Weibull distribution. All covariates reported in Table 3 of the manuscript were included in these models. We compared the fit of these two models using AIC statistics (AIC Gompertz = 24957.55 vs AIC Weibull = 24976.17) and this comparison favored a Gompertz distribution.

We then subjected the Gompertz distribution to a second evaluation. We fit a piecewise exponential model(5) to the data; this type of model splits time into half-year intervals and estimates the hazard within each interval. The baseline hazard estimated from this model was plotted against the baseline hazard estimated from the Gompertz AG model.

Figure A1 compares the Gompertz baseline hazard and the baseline hazard estimated from the piecewise exponential model. The comparison showed close concordance between the two hazards and confirmed our choice of the Gompertz distribution, which was then used in all subsequent multivariable analyses.

Figure A1: Comparison of baseline hazard function for a piecewise exponential distribution and a Gompertz distribution



Sensitivity Analyses

We conducted two sensitivity analyses: re-runs of the main multivariable model (1) using only “serious” complaints; and (2) fitting a PWP-CP model instead the AG model. **Table A2** shows coefficients and cluster-adjusted standard errors from these two re-runs, alongside those from the main multivariable model.

Table A1: Multivariate survival analysis estimating risk of recurrent complaints – main model and two sensitivity analyses

	(A) Main model: AG		(B) Sensitivity analysis: PWP-CP		(C) Sensitivity analysis: Serious complaints only	
	Coef.	Robust SE	Coef.	Robust SE	Coef.	Robust SE
Number of prior complaints						
1 (ref)	--	--	--	--	--	--
2	0.66	0.04	--	--	0.68	0.05
3	1.17	0.06	--	--	1.27	0.08
4	1.51	0.08	--	--	1.69	0.10
5	1.82	0.10	--	--	1.93	0.14
6	2.18	0.11	--	--	2.35	0.18
7	2.26	0.13	--	--	2.42	0.17
8	2.25	0.15	--	--	2.33	0.21
9	2.78	0.16	--	--	2.72	0.25
10 or more	3.39	0.22	--	--	3.70	0.30
States and Territories						
1 (ref)	--	--			--	--
2	0.80	0.09	0.91	0.10	0.81	0.12
3	0.74	0.09	0.81	0.11	0.84	0.13
4	0.64	0.11	0.71	0.12	0.62	0.15
5	0.62	0.10	0.70	0.11	0.58	0.14
6	0.55	0.12	0.64	0.14	0.60	0.17
7	0.22	0.10	0.24	0.12	0.14	0.14
Male doctor	0.31	0.05	0.36	0.06	0.33	0.08
Urban practice location	-0.02	0.04	0.001	0.05	0.01	0.06
Specialty of doctor						
Plastic surgery	0.71	0.08	0.86	0.08	0.92	0.09
Dermatology	0.45	0.09	0.60	0.13	0.59	0.13
Obstetrics and gynaecology	0.41	0.08	0.47	0.12	0.54	0.09
Orthopaedic surgery	0.27	0.05	0.29	0.06	0.33	0.07
Other surgery	0.26	0.05	0.31	0.06	0.36	0.07
General surgery	0.37	0.11	0.39	0.12	0.59	0.13
Ophthalmology	0.18	0.08	0.13	0.09	0.26	0.10
Psychiatry	0.14	0.06	0.18	0.08	0.15	0.08
General practice (ref)	--	--	--	--	--	--
Internal medicine	-0.07	0.08	-0.09	0.10	-0.06	0.11
Radiology	-0.11	0.50	-0.13	0.58	-0.21	0.49
Anaesthesia	-0.43	0.10	-0.48	0.11	-0.57	0.15
Other	-0.44	0.12	-0.49	0.14	-0.39	0.16
Age of doctor						
<35 years	--	--			--	--
36 to 45 years	0.27	0.07	0.32	0.09	0.18	0.10
46 to 55 years	0.34	0.07	0.39	0.08	0.31	0.10
56 to 65 years	0.36	0.08	0.44	0.09	0.31	0.10
Constant	-3.04	0.12	-2.90	0.14	-3.39	0.16

Table A1, continued.

	(A) Main model: AG		**(B) Sensitivity analysis: PWP-CP**		**(C) Sensitivity analysis: Serious complaints only**	
	Coef.	Robust SE	Coef.	Robust SE	Coef.	Robust SE
Gamma	-0.21	0.01	--	--	0.20	0.01
k = 2	--	--	0.18	0.01	--	--
k = 3	--	--	0.25	0.01	--	--
k = 4	--	--	0.27	0.02	--	--
k = 5	--	--	0.31	0.02	--	--
k = 6	--	--	0.35	0.02	--	--
k = 7	--	--	0.35	0.02	--	--
k = 8	--	--	0.34	0.03	--	--
k = 9	--	--	0.39	0.03	--	--
k = 10	--	--	0.45	0.03	--	--
Constant	--	--	-0.30	0.02	--	--

The first sensitivity analysis is described in the manuscript. A full set of results from it are shown in column C of Table A1 (above).

The second sensitivity analysis used the PWP-CP model, which involves relaxation of the common baseline hazard assumption in the AG model. This model is described in the “Choice of statistical model” subsection above. Comparisons of the estimates from the PWP-CP model (column B in Table A1) and main model (column A) suggest that the shape of the baseline hazard differs across the strata (that is, the number of previous complaints). However, this difference appears to have negligible effects on the values of the coefficients and standard errors for the other variables in the model, which are very similar. In sum, the comparison supports the view that our results on predictors such as specialty, doctor sex, doctor age etc. are not sensitive to choice of the AG model over the PWP-CP model.

Model diagnostics

We conducted several tests to evaluate the specification and fit of main multivariable model.

To assess model goodness of fit, we plotted partial Cox-Snell residuals from the main model against the empirical cumulative hazard function (derived from Kaplan-Meier values). The results are shown in **Figure A2**. The residuals follow the straight line closely between values of 0 and 4, but deviate from the line thereafter. The deviant values represent only a small fraction of the data (*n* = 18 complaints out of thousands); the dotted box in the figure shows where 99.9% of the data lies. Thus, the plot suggests that the model provides a reasonable fit for virtually all of the data.

Figure A3 shows the cumulative Cox-Snell residuals plotted against time, a plot that is useful in detecting influential observations. A total of 26 doctors have influence values greater than 10. Re-estimation of the main model without these observations did not alter the parameter estimates, which suggests minimal influence of "outlier" observations on our results.

Figure A2: Cox-Snell residuals to evaluate model fit.

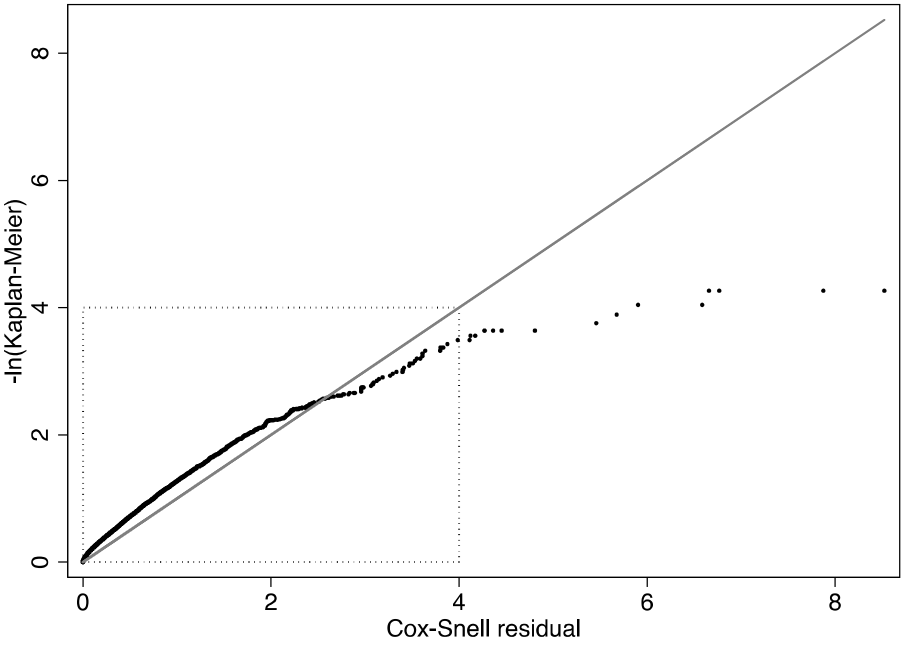
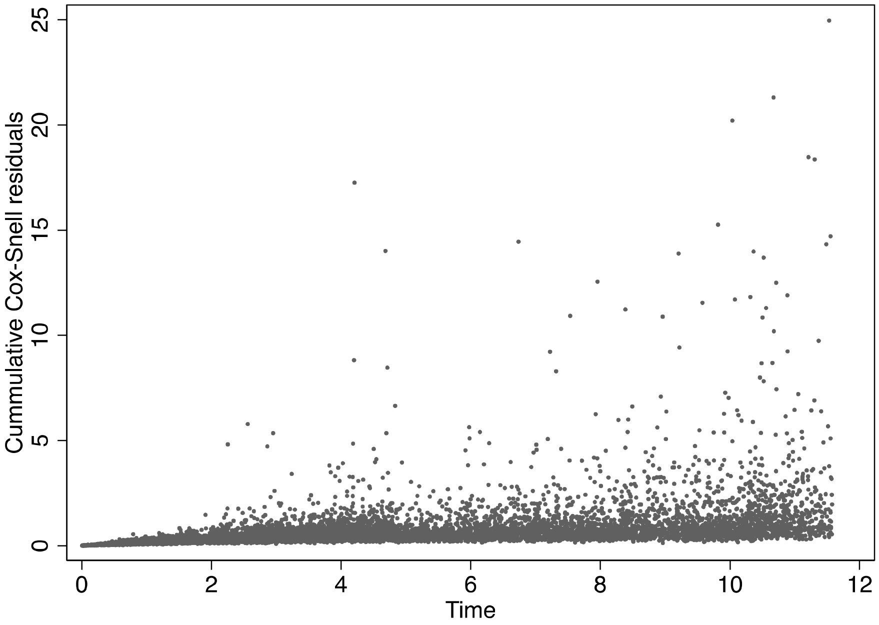


Figure A3: Cumulative Cox-Snell residuals to identify influential observations



Population-level Risk Predictions

Typically, Kaplan-Meier curves are plotted to show differences in survival (or failure) between two or more groups. This approach is also useful in observational studies, such as ours, but it needs to be modified to reflect adjustments made during modeling to accommodate the influence of measured risk factors that potentially confound the association between the exposure and the outcome.

One method of adjustment involves comparing survival between "constructed" subjects with collections of fixed characteristics (e.g. male middle-aged GPs practicing in urban settings vs female middle-aged GPs practicing in urban settings, etc). A second method of adjustment involves examining the effect of a selected variable while setting the values of all other covariates to their mean values. Neither of these approaches is particularly attractive. The fixed construct method complicates presentation of findings and becomes difficult with more than a couple of variables of interest. The averaging method is suspect for certain fixed categorical covariates, such as sex, because mean values for these variables are intrinsically artificial.

A third method of adjustment, sometimes referred to as adjusted survival curves, (6,7) is more attractive and well-suited to our study. In this method, survival curves are estimated for each individual using results from a multivariable analysis, and then averaged at each of a series of closely-spaced time points to plot fitted curves. The averaging assumes everyone takes the same stated level of the variable of interest, but that otherwise their covariate pattern is unchanged. The process is a specific example of Robins' G-computation(8), an approach for estimating the causal effect of an exposure, which Snowden et al(9) have shown is equivalent to a form of model-based standardization where the reference group is the observed study population. The method is flexible and able to handle multiple fixed and time-dependent covariates, and an unbalanced distribution of covariates.(6)

In applying this method to our analysis, we began calculating the adjusted survival curves by computing a failure function for each individual in the study

$$F(t) = 1 - \exp\{-\lambda\gamma^{-1}(e^{\gamma t} - 1)\}$$

where $\lambda_j = \exp(x_j\beta)$ and γ is the ancillary parameter estimated from the data (based on the Gompertz distribution). Using the coefficients from the main model, we calculated $F(t)$ over a grid of values for *t*, ranging from *t* = 0 to *t* = 5 years in 1 month increments.

For example, to isolate the effect of a sex, we calculated survivor functions for all doctors by forcing this variable to have a value of "male", while leaving all other covariates at their observed values. The calculations were done over the grid of time values, and then averaged at each time value. The procedure was then repeated after forcing the variable to have a value "female".

Adjusted failure curves for the "number of previous complaints" variable are reported in the main paper. To calculate these estimates, we recoded all observations so that each individual had experienced only 1 complaint, but all other variables remained at their observed values. We calculated $F(t)$ for each individual and then averaged these values for each value of *t*. This process was repeated, setting number of prior complaints to 2 complaints, 3 complaints, 5 complaints and 10+ complaints to produce the plots shown in Figure 2 (Panel A) of the manuscript.

We calculated analogous estimates for other significant covariates, but have elected not to show them due to limitations of space. However, adjusted survival curves for doctor specialty and doctor sex are shown below.

Figure A4: Specialty of doctor

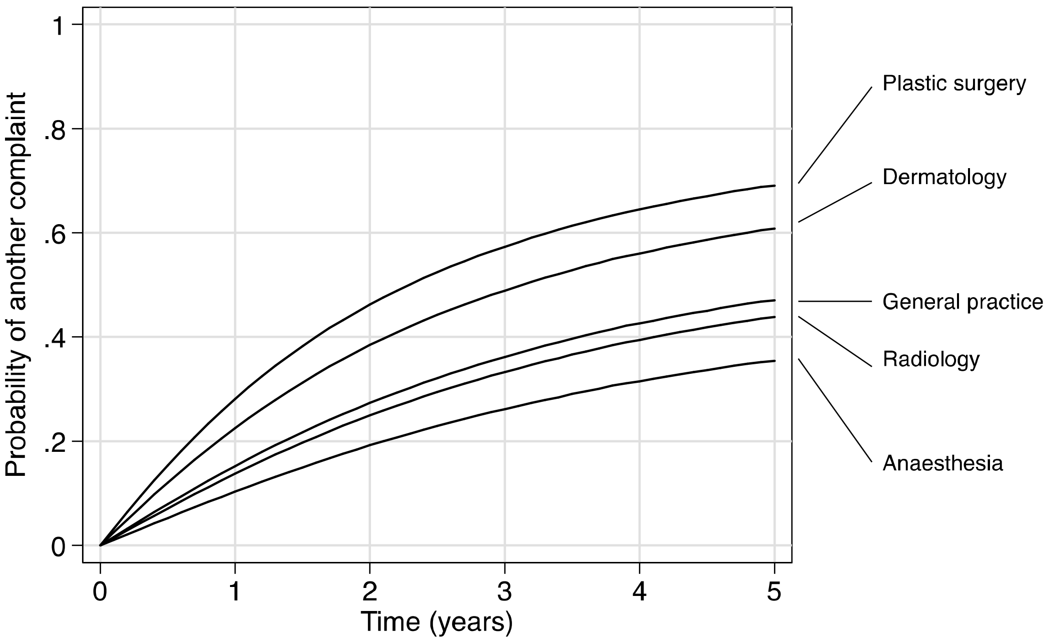
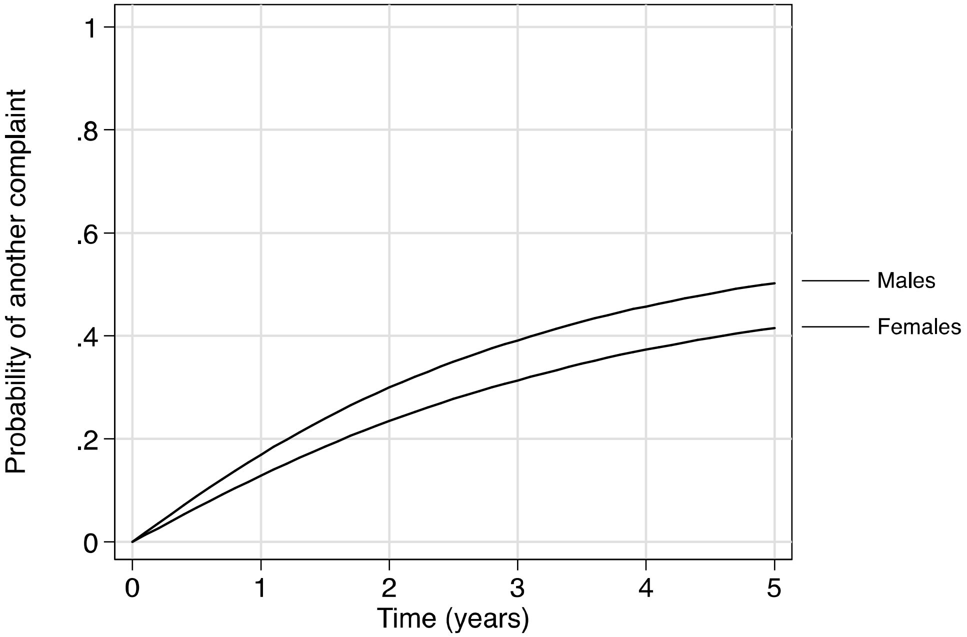


Figure A5: Sex of doctor



EXAMPLES OF COMPLAINTS

Example 1:
A male patient complained that there had been a 12 month delay in diagnosis of his non Hodgkin's Lymphoma. He had attended a general practitioner's clinic 4 times and described consistent symptoms. The GP eventually referred him for an x-ray. The man was told the x-ray was reported as 'normal'. It was only after he was referred for an emergency CT scan that the x-ray was reviewed. He received a financial settlement.

Coding: Clinical care - diagnosis

Example 2:
To treat endometriosis, a female patient had undergone a surgical procedure to divide the ligaments around her uterus. She experienced chronic pain following the procedure, something she claimed the obstetrician-gynaecologist who did the procedure had not discussed with her. The obstetrician-gynaecologist did not deny this. His response letter stated: "I did not really discuss the risks of the procedure. The husband was sick of having a wife who was always in pain, so really there was no choice."

Coding: Communication - consent

References

1. Australian Health Practitioner Regulation Agency - Registers of Practitioners [Internet]. AHPRA. [accessed Feb. 2012]. Available from: http://www.ahpra.gov.au/Registration/Registers-of-Practitioners.aspx

2. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics. 1982;10(4):1100–1120.

3. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1980;68(2):373–379.

4. Hosmer DW, Lemeshow S, May S. Applied survival analysis. Wiley; 2008.

5. Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. Stata Press; 2011.

6. Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new method. Am J Epidemiol. 1996;143(10):1059–1068.

7. Makuch RW. Adjusted survival curve estimation using covariates. Journal of chronic diseases. 1982;35(6):437–443.

8. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986.

9. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol. 2011;173(7):731–738.