Estimating the effect of health service delivery interventions on patient length of stay

Watson, Samuel I.; Lilford, Richard J.; Sun, Jianxia; Bion, Julian

DOI:
10.1111/rssc.12501

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Watson, SI, Lilford, RJ, Sun, J & Bion, J 2021, 'Estimating the effect of health service delivery interventions on patient length of stay: a Bayesian survival analysis approach', *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 70, no. 5, pp. 1164-1186. https://doi.org/10.1111/rssc.12501

Received: 10 March 2020 | Accepted: 27 April 2021

DOI: 10.1111/rssc.12501

ROYAL STATISTICAL SOCIETY
Series C Applied Statistics

ORIGINAL ARTICLE

Estimating the effect of health service delivery interventions on patient length of stay: A Bayesian survival analysis approach

Samuel I. Watson[1] | **Richard J. Lilford**[1] | **Jianxia Sun**[2] | **Julian Bion**[1]

[1]University of Birmingham, Birmingham, UK

[2]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Correspondence
Samuel I. Watson PhD, Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom.
Email: s.i.watson@bham.ac.uk

Funding information
National Institute for Health Research; Health Services and Delivery Research Programme, Grant/Award Number: 12/128/17; NIHR Applied Research Centre WM

Abstract

Health service delivery interventions include a range of hospital 'quality improvement' initiatives and broader health system policies. These interventions act through multiple causal pathways to affect patient outcomes and they present distinct challenges for evaluation. In this article, we propose an empirical approach to estimating the effect of service delivery interventions on patient length of stay considering three principle issues: (i) informative censoring of discharge times due to mortality; (ii) post-treatment selection bias if the intervention affects patient admission probabilities; and (iii) decomposition into direct and indirect pathways mediated by quality. We propose a Bayesian structural survival model framework in which results from a subsample in which required assumptions hold, including conditional independence of the intervention, can be applied to the whole sample. We evaluate a policy of increasing specialist intensity in hospitals at the weekend in England and Wales to inform a cost-minimisation analysis. Using data on adverse events from a case note review, we compare various specifications of a structural model that allows for observations of hospital quality. We find that the policy was not implemented as intended but would have likely been cost saving, that this

conclusion is sensitive to model specification, and that the direct effect accounts for almost all of the total effect rather than any improvement in hospital quality.

KEYWORDS

Bayesian, causal model, direct and indirect effects, health services research, survival analysis

1 | INTRODUCTION

Health care interventions can be classified according to where they are located in the causal pathway between health care policy and clinical outcomes (Lilford et al., 2010). Clinical interventions, such as pharmaceutical and medical products, act directly on the patient to cause physiological changes resulting in clinical outcomes. At one stage removed, interventions may target how a clinical product is applied, for example checklists or care guidelines. *Service delivery interventions* act yet further 'upstream' from the patient and target clinical or higher level processes such as staffing decisions and system organisation. Service delivery interventions typically have more diffuse effects acting through multiple causal pathways to affect patient outcomes, which can present challenges to estimating their effects. For example, service delivery interventions can cause changes to which patients are treated in the hospital, and their effects are often mediated by hospital 'quality', which is confounded with patient health status. We define 'quality' in the technical sense as (the avoidance of) preventable harm to patients through errors in care, such as diagnostic, medical, or surgical errors (Brennan et al., 1991).

The example of a service delivery intervention we examine is the 'seven day services' policy launched by the UK government in 2013 (Department of Health, 2015). There were a number of studies prior to 2015 that showed an increased risk of mortality among patients admitted on a weekend when compared to a weekday, the so-called 'weekend effect' (e.g. Bell & Redelmeier, 2001; Freemantle et al., 2012, 2015) We conducted a meta-analysis of 68 studies of the weekend effect and estimated the pooled odds of mortality among weekend admissions from these studies to be 16% higher (Chen et al., 2019). One hypothesis to explain this evidence was that care quality was lower at the weekend, particularly due to the absence of specialists (Bell & Redelmeier, 2001). Indeed, we showed that there was lower specialist intensity at the weekend in hospitals in England and Wales in terms of number of specialist hours per emergency admission (Aldridge et al., 2016). However, a growing body of evidence provides an alternative explanation for the weekend effect. Weekend admissions are sicker on average than their weekday counterparts (Mohammed et al., 2017; Sun et al., 2019), and in addition, fewer patients are admitted despite similar emergency department attendance rates (Meacock et al., 2017). A key aspect of the 7-day services policy was to encourage hospitals to increase the ratio of specialists to patients ('specialist intensity') at the weekend, to weekday levels (Department of Health, 2015; NHS England, 2013). In this article, we propose an empirical approach to inference of the association between length of stay and service delivery interventions. We estimate the relationship between specialist presence and patient length of stay to inform a simple cost-minimisation analysis of a policy of increasing specialist intensity. Length of stay is the main driver of costs in the healthcare system; it is not only related to quality of care as sicker patients stay longer, but also sensitive to other clinician functions, such as patient discharge as we describe below, so it makes for an ideal outcome for these analyses.

Figure 1 shows a causal model, which illustrates our assumptions about the general causal mechanisms of action of hospital quality improvement and service delivery interventions. In this set up we have patient length of stay T, mortality M, the intervention I, the patient's health status at initial presentation H, hospital quality Q and an indicator for whether the patient has been admitted A. The intervention has a direct effect by affecting timely discharge, as specialists are responsible for discharging patients, an indirect effect through Q by changing the risk of experiencing preventable harm, and an effect on the probability of admission, which is itself a function of patient health. We assume that patient health affects the risk of preventable harm as sicker patients require a greater number of, and more complex, treatments. Finally, length of stay and the risk of mortality are functions of both quality and patient health. An implication of this model is that a service delivery intervention can both reduce average length of stay by improving hospital quality, and increase it if less severely ill patients are no longer admitted, or patients survive who would have otherwise died.

1.1 | Survival analysis

There is a large literature devoted to modelling patient length of stay in hospital. One of the main challenges is that discharge time is not observed for patients who die. A standard survival analysis model allows for the right censoring of discharge times, where the censoring mechanism is often assumed to be uninformative. However, ignoring the dependence between length of stay and the risk of mortality may lead to biased estimates of intervention effects. An approach often used to jointly model time-to-event and other outcomes is a 'shared parameter' or 'shared frailty' formulation: conditional on some random effect(s) α, which represents health H in our causal model (Figure 1), the joint model can be factorised as $p(T, M, \alpha)=p(T|\alpha)p(M|\alpha)p(\alpha)$, where $p(.)$ is a probability density function. This approach has been widely used for jointly modelling longitudinal outcomes and time-to-event data (e.g. Graham et al., 2011; Henderson et al., 2000; Long & Mills, 2018; Wang & Taylor, 2001; Zhang et al., 2010, 2017). More complex models, including for semi-competing risks (in which a non-terminal outcome is censored by a terminal outcome), also adopt variations of this formulation (e.g., Cook & Lawless, 1997; Lee et al., 2015; Zeng & Lin, 2009) as it permits flexible dependence structures and can be extended to more complex structural models. Both parametric and semi-parametric baseline hazard models have been assumed in the literature for the survival component of the model. While much of the literature focuses on frequentist inference for these models, Bayesian approaches are useful here given their flexibility in handling complex random effect structures where maximum likelihood based methods may be precluded, and their ability to propagate uncertainty through structural models, particularly given the interest in effect decomposition described below.

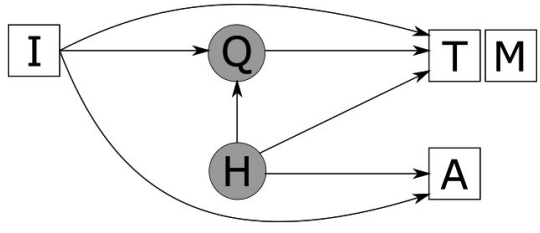

FIGURE 1 Directed acyclic graph indicating assumed causal relationships with squares representing observed variables and shaded circles unobserved variables. I = intervention, Q = quality, H = health, A = admission, T = length of stay, M = mortality

1.2 | Post-treatment selection bias

Post-treatment selection bias arises from only using the observations of admitted patients and so implicitly conditioning on *A*. Many service delivery interventions will affect the health threshold at which patients are admitted so the sample of patients under the treatment and control condition will differ on both observable and unobservable variables. Indeed, this was one of the principle arguments for the existence of the weekend effect (Meacock et al., 2017). Moreover, the selection bias would exist whether or not the treatment was randomised to clusters. There are a number of articles that discuss the bias caused by conditioning on post-treatment confounders (e.g. Montgomery et al., 2018; Rosenbaum, 1984). However, this issue is rarely addressed in evaluations of service delivery interventions even where they are attentive to the issue of non-random treatment allocation. An instrumental variable approach has been widely used in evaluations to estimate the 'causal effect' of attending a particular hospital, which can then be compared between hospitals by intervention status (e.g. Gaynor et al., 2016; Gowrisankaran & Town, 1999; Watson et al., 2017). Conceptually, this 'randomises' patients to hospitals and permits identification of local average treatment effects. However, if the admissions process is not independent of the intervention then which of the patients 'randomised' to each hospital are admitted would change by intervention status and patient health would not be comparable by intervention status. For example, Gaynor et al. (2016) evaluated a 2006 policy in England of expanding patient choice over the hospital they attend and used the nearest hospital as an instrument for the treating hospital to estimate the causal effect of the policy on the risk of mortality in each institution. However, the policy may have affected which patients are admitted because of changes to the flows of patients between hospitals, leading to the problem of post-treatment selection bias. Watson et al. (2017) used the same empirical strategy to examine the relationship between staffing levels in neonatal intensive care units and patient outcomes, but better staffed hospitals may admit different cadres of patients.

Since the conditioning is implicit due to observing only admitted patients, it cannot simply be 'undone' by removing the conditioning. As a possible example of specialist presence affecting patient admissions, Jena et al. (2015) examined patients admitted with acute myocardial infarction during national cardiology meetings when cardiologists were not present and found evidence 30-day mortality was *lower* than during normal times, which they suggested may be explained by the reduced throughput of patients needing more complex treatment. In the context of the weekend effect literature, Meacock et al. (2019) used data from both attendances and admissions to model the selection (i.e., admissions) mechanism using a Heckman selection model approach and found a significantly reduced, although not eliminated, estimate of the weekend effect.

In the absence of observations on attending patients, an alternative empirical strategy is to limit analyses to a group of patients for whom the probability of admission is (conditionally) independent of the intervention. One example of this may be patients requiring intensive care who would all likely be admitted regardless of the presence of any intervention. For example, Kerlin et al. (2013) compared night-time specialist staffing to no staffing at nights in an ICU in a randomised trial. Different night shifts were randomly assigned to have specialist staff or not. Little evidence was found of changes to length of stay by the presence of staff on day of admission. But this does not take into account staffing more broadly in the hospital, either in the emergency department or on the units on which patients sojourn after leaving the ICU. Indeed a large number of studies in this area have been focused on intensive care units, likely due to the quality and availability of relevant data. Generally though, they do report a lower risk of mortality with higher levels of staffing (Galloway et al., 2018; Wilcox et al., 2013).

1.3 | Direct and indirect effects

Service delivery interventions can have both direct and indirect (acting through quality) effects on patient clinical outcomes. The decomposition of these effects can have important implications for policy. For example, if the predominant effect of the intervention is not due to its effects on quality but by other means, such as improving timely discharges, then a less costly alternative may achieve the same results. Typically both quality and health are unobserved, which prevents decomposition of intervention effects into direct and indirect components. Importantly though, even if the decomposition of effects was not itself of primary interest, the total effect of the intervention would be confounded if health H were not fully observed due to the relationship between health and quality.

The assumption that patient health can be fully observed is often used in quantitative evaluations of hospitals, despite little evidence to support it. Under the assumption that we can fully adjust for patient health then it is assumed that any remaining variation in mortality rates, or other outcome, between hospitals is attributable to differences in quality. This is the basis of the standardised mortality ratio (SMR) (Mohammed et al., 2009). However, the SMR has been strongly criticised as a poor proxy for preventable mortality. For example, Girling et al. (2012) estimated that if 6% of hospital deaths are preventable (following Hogan et al., 2012) then the predictive value of the SMR for the preventable death rate can be no greater than 9%. Empirical comparisons of estimates of preventable mortality and SMRs in the same institutions support this conclusion and show a low level of correlation (Hogan et al., 2015).

Observations of quality, such as through assessment of preventable errors and harm, are required for reliable estimates of total and decomposed effects, under the causal assumptions of Figure 1. Recently, there has been several approaches proposed for the decomposition of direct and indirect effects via a mediating variable based on Pearl's 'front-door estimator' (Fulcher et al., 2020; Pearl, 2012). One of the key identifying assumptions for this method is that the mediator is independent of any unobserved confounding variables, which, if health is not fully observed, is therefore not met even if quality is observed. However, if it can be assumed that the treatment is (conditionally) independent of patient health then a structural approach should identify both the direct and indirect effects.

2 | DATA

This analysis forms part of the HiSLAC project, which investigated the effect of a roll-out of seven-day services in England and Wales (www.hislac.org). A number of different data sources were compiled for the quantitative analyses in this project. Specialist intensity was assessed by an annual point prevalence survey in each of the 5 years of the project (2014–2018) (Aldridge et al., 2016). Specialists working in participating NHS Trusts in England were invited to take part in a survey on specialist presence and number of hours provided on a specific Wednesday and Sunday in each year. The mean number of emergency admissions for each hospital on Wednesdays and Sundays in those years was derived from the Hospital Episodes Statistics database. From these data sources, specialist intensity was estimated as the total specialist hours per 10 emergency admissions (EAs). The level of specialist staffing we assign to each patient is that recorded for the day (week or weekend) of admission in the admitting hospital recorded by the annual point prevalence survey. This follows the general focus in the hospital quality literature. The majority of patients are discharged within 2 days (see below), and many of the key events, including initiating patients on care pathways, occur at the beginning of a hospital stay, so this convention is somewhat justified. However, we note that the focus on the day

of admission means that what happens on subsequent days of an episode of care is left as unobserved and potentially confounding factors.

To measure hospital quality, data on preventable errors and adverse events were collected by a case record review of 4,000 non-operative admissions, a protocol for which is published elsewhere (Bion et al., 2017). The admissions were randomly sampled from the complete list of admitted patients obtained from the Hospital Episode Statistics (HES) database from 20 selected hospitals in equal number, with 50% weekend and 50% weekday admissions, and 50% in the pre-policy epoch (2014/15) and 50% in the post-policy epoch (2017/8). Records were randomly sampled from all admissions for the complete financial year. Hospitals were purposively selected on the basis of their pre-policy level of Sunday specialist intensity from the 2014 point prevalence survey in order to provide a high level of variance in specialist intensity both between hospitals and over time. Hospitals were stratified into quintiles by bed numbers and then within each quintile the top and bottom two hospitals by Sunday specialist intensity were selected.

The case records were reviewed by 79 clinical experts (consultants and senior registrars). Case records were anonymised by masking the name, address, age, and postcode of the patient, clinician identifiers and by censoring lengths of stay over 7 days. Errors and adverse events were characterised using a structured judgement review, based on a previous typology (Hogan et al., 2015). Reviewers judged the number of errors in each of nine categories, whether they caused harm (i.e. an adverse event), and their preventability. 800 records were reviewed twice to assess reviewer reliability: we randomly selected one of the two reviews from duplicated records for this analysis. There were 37 records where the reviewers could not determine whether an error had occurred due to insufficient information, leaving 3,963 case reviews. Case note review data were matched to patient records from the Hospital Episode Statistics database to extract: patient age, sex, postcode, mode of arrival, day and year of admission, in-hospital death and length of hospital stay. A full summary of the case note review data will be published elsewhere (under review). For the purposes of this study we 'observe' hospital quality using the indicator for whether the patient experienced preventable harm—an adverse event. While data on a range of errors are available, the risk of experiencing an error is independent of patient outcomes conditional on adverse events. Given the complicated nature of error classification, particularly errors without consequence, we opt to use only adverse events and not the broader error data. Agreement between reviewers for errors and adverse events was low (intraclass correlation coefficients ranged from 0.003 to 0.131), which reflects the poor agreement seen in other case note reviews (Hogan et al., 2015). However, inference is based on a model of hospital-level mean effects, and the intervention also varies at this level. Aggregated to the hospital level agreement on mean hospital quality was much better (0.138 to 0.883).

3 | STATISTICAL MODEL

3.1 | Objectives

In what follows, let the time to discharge alive (length of stay) for patient $i = 1, \ldots, N$ in hospital $j = 1, \ldots, J$ be T^*_{ij} and the quality indicator is Q_{ij}. The objective of our analysis is to determine the average absolute difference in length of stay between two levels of specialist intensity $I = l$ and $I = l^*$, say, and decompose this effect into direct and indirect pathways. Let $T^*_{ij}(l, Q_{ij}(l))$ denote the counterfactual length of stay if specialist intensity had taken the value l acting both directly and indirectly through Q. Then the average treatment effect is

$$\begin{aligned}\text{ATE} &= E\left[T^*_{ij}(l,\, Q_{ij}(l))\right] - E\left[T^*_{ij}(l^*,\, Q_{ij}(l^*))\right] \\ &= E\left[T^*_{ij}(l,\, Q_{ij}(l)) - T^*_{ij}(l,\, Q_{ij}(l^*))\right] + E\left[T^*_{ij}(l,\, Q_{ij}(l^*)) - T^*_{ij}(l^*,\, Q_{ij}(l^*))\right]\end{aligned} \tag{1}$$

where the second line is the 'natural' decomposition into indirect and direct effects, respectively (Fulcher et al., 2020).

Estimation of the effect of service delivery interventions may not be possible using only data from admissions if the intervention affects admission probability. The empirical strategy we consider here is to conduct our primary analysis on a subsample of the data for which we assume $I \perp\!\!\!\perp A$ (Figure 1), that is the admissions mechanism is plausibly independent of the level of specialist intensity. We select patients who arrive at the hospital by ambulance, who we assume the majority of are admitted and previous studies have suggested low variation in admission propensity for arrivals by ambulance by day of week (Anselmi et al., 2017). However, there remains the question of the generalisability of results from this subsample to the broader patient population. For a given patient, the various counterfactual expected lengths of stay can be related by a set of parameters, π_1, π_2, and π_3:

$$\begin{aligned}E\left[T^*_{ij}(l,\, Q_{ij}(l))\right] &= \pi_1 E\left[T^*_{ij}(l^*,\, Q_{ij}(l))\right] \\ E\left[T^*_{ij}(l,\, Q_{ij}(l))\right] &= \pi_2 E\left[T^*_{ij}(l,\, Q_{ij}(l^*))\right] \\ E\left[T^*_{ij}(l,\, Q_{ij}(l))\right] &= \pi_3 E\left[T^*_{ij}(l^*,\, Q_{ij}(l^*))\right] = \pi_1\pi_2 E\left[T^*_{ij}(l^*,\, Q_{ij}(l^*))\right]\end{aligned} \tag{2}$$

Under a proportional hazards assumption (Section 3.2) the parameters π_1, π_2 and π_3 would be the same for different subsamples and equivalent to an inverse hazard ratio. Differences in estimates of these parameters from different subsamples implies either a failure of the proportional hazards assumption or of the assumption that $I \perp\!\!\!\perp A$ in at least one of the subsamples. Figure 2 shows the Kaplan–Meier estimates of the survival function and log cumulative hazard by mode of arrival. If the proportional hazards assumption holds then the survival functions should not cross over between groups and the log cumulative hazard

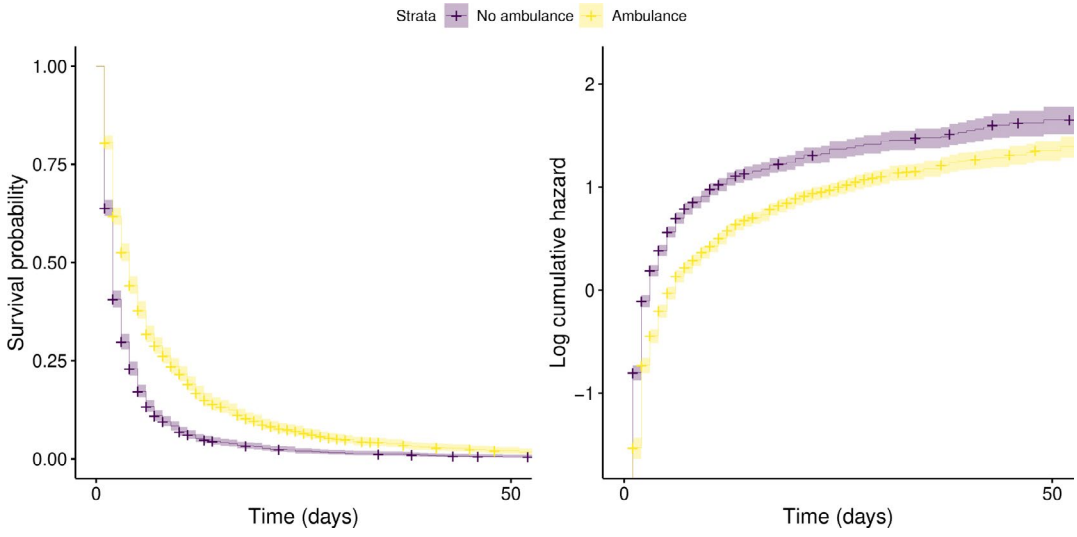

FIGURE 2 Kaplan–Meier estimates of the survival function (left) and log cumulative hazard (right) with 95% confidence intervals by mode of arrival

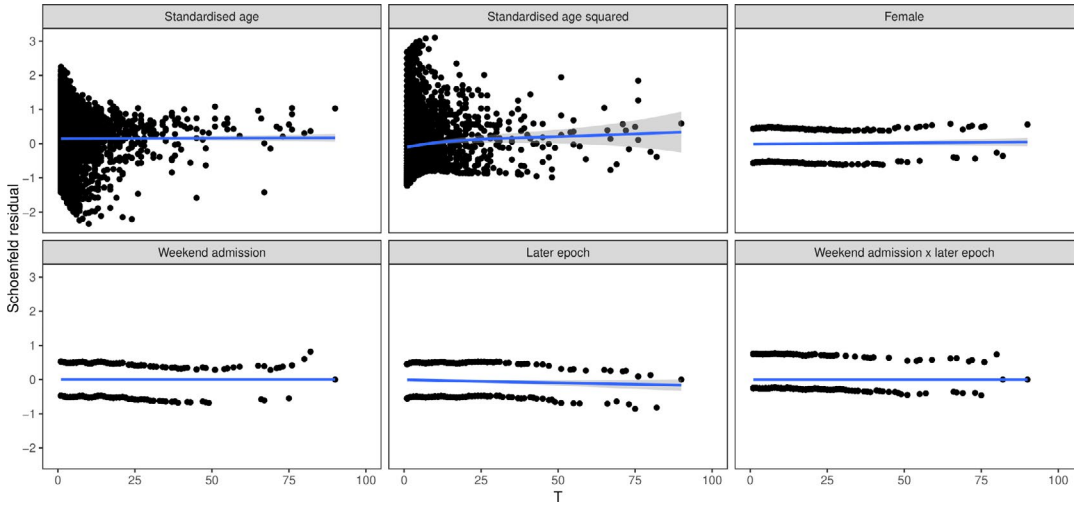

FIGURE 3 Schoenfeld residuals for each covariate from the preferred model specification (see Sections 3 and 4) evaluated at the posterior mean of the model parameters with smoothed trend line

functions should be parallel (Hess, 1995). As an additional check of the proportional hazards assumption, we estimate the Schoenfeld residuals for each covariate in the model from the preferred model described in subsequent sections, evaluated at the posterior mean of the model parameters. Under a proportional hazards assumption, the Schoenfeld residuals are independent of time (Schoenfeld, 1982). Figure 3 shows these residuals with a smoothed trend line and there is no relationship evident. The proportional hazards assumption therefore appears reasonable, and so we attribute any differences between estimates of π_1 or π_2 between the subsample and the whole sample, to a failure of the independence assumption in the broader sample. We use estimates of the hazard ratio from the primary subsample for the cost-minimisation analysis.

3.2 | Bayesian survival model

Following a traditional survival analysis setup, we do not observe T_{ij}^* for all patients, only for those who do not die. Our data are $\mathscr{D} = \left[Y_{ij}\right] = \left[T_{ij}, M_{ij}, Q_{ij}\right]$ where T_{ij} is the time the patient was discharged or died, and M_{ij} is a dichotomous indicator for whether the patient died. We also observe Q_{ij}, a dichotomous indicator for whether the patient experienced at least one adverse event (i.e. preventable harm). Figure 1 represents our underlying causal assumptions for the model; we assume that the mortality mechanism and discharge time are independent conditional on health, quality and the intervention. We assume that health can be captured by the hospital level random effect α_j and observed covariates $\mathbf{x}_{ij}$ conditional on which we can factorise the joint density of T_{ij}, M_{ij} and Q_{ij} following a 'shared parameter' type formulation:

$$f(Y_{ij}, \alpha_j \,|\, \mathbf{x}_{ij}, \boldsymbol{\theta}) = f(T_{ij} \,|\, \mathbf{x}_{ij}, \alpha_j, \boldsymbol{\theta}, Q_{ij}) f(M_{ij} \,|\, \mathbf{x}_{ij}, \alpha_j, \boldsymbol{\theta}, Q_{ij}) f(Q_{ij} \,|\, \mathbf{x}_{ij}, \alpha_j, \boldsymbol{\theta}) f(\alpha_j \,|\, \mathbf{x}_{ij}, \boldsymbol{\theta}) \tag{3}$$

with model parameters $\boldsymbol{\theta}$. To specify each of the components of the model, we first turn to length of stay. The density function is the product of hazard and survival functions, $f(T_{ij} \,|\, \mathbf{x}_{ij}, \alpha_j, \boldsymbol{\theta}, Q_{ij}) = f_{ij}(t) = h_{ij}(t) S_{ij}(t)$ where $S_{ij}(t) = \exp(-\int_0^t h_{ij}(u)\, du)$. We specify the following proportional hazards model:

$$h_{ij}(t) = h_0(t)\exp(\eta^1_{ij}) \tag{4}$$

where $h_0(t)$ is the baseline hazard function and the linear predictor is:

$$\eta^1_{ij} = \mathbf{x}'_{ij}\beta^1 + \alpha_j + f_1(I_j) + \delta^1 Q_{ij} \tag{5}$$

where $f(I_j)$ is a smooth function of specialist intensity, $\mathbf{x}_{ij}$ is a vector of patient-level covariates described below, $\alpha_j \sim N(0,\ \sigma^2_\alpha)$ is the hospital-level random effect, and β^1 and δ are model parameters. For the function f_1, we compare linear, quadratic and cubic polynomial specifications. We also consider three different specifications of the baseline hazard $h_0(t)$:

- Exponential: $h_0(t) = \lambda_1$.
- Weibull: $h_0(t) = \lambda_2 t^{\lambda_2 - 1}$ where $\lambda_2 > 0$ is the shape parameter.
- Semi-parametric: We specify $h_0(t) = \sum_{b=1}^{B} \lambda_{3b} M_b(t; d = 3,\ k)$ where $M_b(t;\ d = 3,\ k)$ is the bth basis term of a degree 3 M-spline with knot locations k, which we set at the upper and low tertiles, and λ_{3b} are M-spline coefficients. We include an intercept and constrain the coefficients to a simplex to ensure identifiability of the coefficients and the intercept in the linear predictor (Brilleman et al., 2020).

Mortality is modelled as a Bernoulli-logistic regression:

$$\begin{aligned} M_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ p_{ij} &= \frac{1}{1+\exp(-\eta^2_{ij})} \\ \eta^2_{ij} &= \mathbf{x}'_{ij}\beta^2 + \psi^2\alpha_j + f_2(I_j) + \delta^2 Q_{ij} \end{aligned} \tag{6}$$

where ψ^2 is a factor loading and the other terms are as for η^1_{ij}. And we similarly specify for the adverse event outcome:

$$\begin{aligned} Q_{ij} &\sim \text{Bernoulli}(q_{ij}) \\ q_{ij} &= \frac{1}{1+\exp(-\eta^3_{ij})} \\ \eta^3_{ij} &= \mathbf{x}'_{ij}\beta^3 + \psi^3\alpha_j + f_3(I_j). \end{aligned} \tag{7}$$

In addition to using data from different subsamples, we compare of estimates of the effect of changes to specialist intensity from this full model to a 'reduced' model in which quality is not observed. For this, we remove terms with Q_{ij} from the linear predictors and the model for Q_{ij} in Equation (7).

3.3 | Patient-level covariates

The observable characteristics of each patient that are used as covariates in each model were chosen based on availability and completeness as well as previous analyses in this area (e.g., Anselmi et al., 2017; Gaynor et al., 2016; Gowrisankaran & Town, 1999; Meacock et al., 2017, 2019; Watson et al., 2017). They are age, age squared, and sex. In addition we included indicators for day of admission (weekend/weekday), epoch and their interaction.

3.4 | Inference

3.4.1 | Prior distributions

We specify weakly informative prior distributions for model parameters. While there exists some prior evidence on the risk of experiencing preventable harm and patient length of stay that may support informative prior distributions, it is highly variable and provides little information to support the location or scale of model parameters relating to specialist intensity or in this particular survival model setting. Weakly informative prior distributions provide a degree of regularisation and facilitate computation compared to 'non-informative' priors, which place a large amount of the density on extreme values of model parameters (Gelman et al., 2008), but provide little information about parameter values inside of a plausible range and therefore have little effect on resulting model estimates. On this basis all model parameters in the linear predictor are assigned $N(0, 5^2)$ priors, and we also set $\sigma_\alpha \sim N(0, 1)[0, \infty)$ and $\lambda_2 \sim N(0, 1)[0, \infty)$. This, for example, sets a prior 95% credible interval on hazard ratios of $[5.5\times10^{-5}, 1.8\times10^{4}]$.

3.4.2 | Log-likelihood

The log-likelihood is given by

$$\begin{aligned}\log L(Y_{ij}|\boldsymbol{\theta}, \alpha_j) = & M_{ij}\left[\log(p_{ij}) + \log(S_{ij}(t))\right] + (1 - M_{ij})\left[\log(1 - p_{ij}) + \log(f_{ij}(t))\right] \\ & + Q_{ij}\log(q_{ij}) + (1 - Q_{ij})\log(1 - q_{ij})\end{aligned}$$

which takes into account the right-censoring due to mortality. We do not observe the length of stay for patients who die, and so they only contribute information about the survival function, risk of mortality and hospital quality.

3.4.3 | Estimation

All models were estimated in R 4.0.1 using Stan 2.24 using the R package `cmdstanr`. Model code is provided in the Supplementary Information. We ran each model for 2000 warmup and 2000 sampling iterations across 8 MCMC chains with the aim of achieving a minimum effective sample size of 1000 for each parameter of interest. Convergence was assessed using the R-hat statistic and graphically.

3.5 | Model comparison

We use a number of diagnostic tools to compare between the models with different hazard functions and specifications of specialist intensity. We use the widely applicable information criterion (WAIC) and a leave-one-out cross validation score (LOO-CV) (Vehtari et al., 2017) to compare between models with the same outcomes and data. The WAIC is based on the log pointwise predictive density with a correction factor for the number of parameters to avoid overfitting. The log pointwise predictive density is evaluated over the posterior distribution of the parameters $\boldsymbol{\theta}$:

$$\text{LPPD} = \sum_{i=1}^{n} \log \int f(Y_{ij} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathscr{D})\, d\theta$$

the penalty term is

$$\text{pen}_{WAIC} = \sum_{i=1}^{n} Var(log(Y_{ij} \mid \boldsymbol{\theta}))$$

then the WAIC is

$$\text{WAIC} = -2\text{LPPD} + 2\text{pen}_{WAIC}$$

where lower values indicate superior model fit. The LOO-CV score is based on the conditional predictive ordinate for each observation, which is the probability of observing the data point based on the model estimated using data without that observation:

$$\text{LOO-CV} = -\sum_{i=1}^{n} \log(f(Y_{ij} \mid \mathscr{D}_{-i}))$$

To compare models more broadly, we conduct a series of graphical posterior predictive model checks. For each model, we sample from the posterior predictive distributions of the survival function:

$$p(\widehat{S}_{ij}(t) \mid \mathscr{D}) = \int\int p(\widehat{S}_{ij}(t) \mid \boldsymbol{\theta}, \alpha_j) p(\boldsymbol{\theta}, \alpha_j \mid \mathscr{D}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\alpha_j \tag{8}$$

and discharge times:

$$p(\widehat{T}_{ij} \mid \mathscr{D}) = \int\int p(\widehat{S}_{ij}(t) \mid \boldsymbol{\theta}, \alpha_j) p(\boldsymbol{\theta}, \alpha_j \mid \mathscr{D}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\alpha_j \tag{9}$$

draws from these posterior predictive distributions are compared to the empirical distributions of the observed data.

3.6 | Treatment effects

We evaluate an increase in specialist intensity by estimating the effect of changing from the weekend average specialist intensity ($I = l$) to the weekday average ($I = l^*$). To estimate the ATE in Equation (1), we set $E\left[T^*_{ij}(l, Q_{ij}(l))\right]$ as the mean length of stay at the weekend in the earlier epoch, which is 4.7 days (Table 1). The indirect and direct effects can be estimated from the proportionate change in length of stay given by the models respectively as the inverse hazard ratios:

$$\begin{aligned} \pi_1 &= \frac{1}{\exp[f_1(l) - f_1(l^*)]} \\ \pi_2 &= \frac{1}{\exp[\delta_1(\Pr(Q \mid I = l) - \Pr(Q \mid I = l^*))]} \end{aligned} \tag{10}$$

3.7 | Cost minimisation

Finally, we convert the ATE into an estimate of net costs. A negative expected net cost provides evidence that the policy will reduce costs overall, and if one assumes that it would have no effect on or improve patient health, then a policy maker should implement the policy as it is 'dominant' in an economic sense. Based on the most recent publication of NHS Reference Costs (2017/18) one bed day is set at £373, which is the mean cost of a non-elective excess bed day (NHS England, 2019). The

TABLE 1 Summary statistics of the sample by epoch and day of admission

		Epoch 1		Epoch 2			
	Whole sample	**Weekday**	**Weekend**	**Weekday**	**Weekend**	**Ambulance**	**Not ambulance**
Hospital-level							
N	20	20	20	20	20	20	20
Specialists (hrs/10 EAs)	38.4 (21.5)	52.7 (22.0)	25.5 (15.0)	48.4 (20.7)	26.9 (13.2)	38.4 (21.5)	38.4 (21.5)
Patient-level							
n	3969	993	996	992	988	1882	2085
Age	61.1 (22.3)	60.5 (22.5)	60.7 (22.2)	62.8 (22.2)	60.4 (22.3)	67.4 (21.0)	55.4 (22.0)
Female (%)	53.3	51.5	53.8	52.4	55.6	53.1	53.4
Arrival by ambulance (%)	47.4	44.4	56.0	40.0	49.3	100	0
Adverse event (%)	2.7	3.5	3.2	1.5	2.2	4.1	1.5
In-hospital mortality (%)		4.2	4.9	3.6	4.1	6.7	2.0
Length of stay, median [IQR]	2 [0, 5]	2 [0, 5]	2 [1, 5]	1 [0, 4]	2 [0, 5]	3 [1, 7]	1 [0, 3]
Length of stay, mean (sd)	4.4 (8.6)	4.5 (7.9)	4.7 (8.4)	3.6 (8.7)	4.7 (9.5)	6.1 (10.6)	2.8 (5.9)

direct per-patient cost of achieving a specialist intensity $I = l$ is based on the pay scales (approximately £80,000 to 110,000), clinical excellence award levels (approximately £3000 to 80,000), and superannuation costs (21%) for specialists in NHS England (British Medical Association, 2019). There is no information in the public domain about the distribution of specialists across pay scales or clinical excellence awards. We therefore take the top end of the pay scale (with superannuation costs) of £77 as the expected cost per specialist hour for the cost-minimisation analysis.

4 | RESULTS

4.1 | Summary statistics

Summary statistics for the sample are reported in Table 1. Generally, there was little change in the patient population when considering age or sex by weekday or weekend admission between epochs. There was some decrease in the proportion of patients arriving by ambulance over time but this occurred on both weekends and weekdays. Similarly there was a reduction in the overall proportion of patients experiencing adverse events for both weekend and weekday admissions. Indeed there appears to be little evidence of a 'weekend effect' in either epoch for variables other than mortality. Specialist intensity also declined on average over weekdays. It is likely that overall patient volumes increased between the two epochs which had the effects of (i) reducing specialist intensity ratios; (ii) increasing the volume of less severely ill patients and as a result (iii) reducing the average risk of an adverse event. Patients who arrived by ambulance were on average older than those who did not arrive by ambulance, they were more likely to both experience an adverse event and to die, and had longer lengths of stay.

Figure 4 shows the levels of specialist intensity at the participating sites on weekends and weekdays in the two epochs. While the majority of sites were estimated to have increased their level of specialist presence at the weekend, most also reduced the specialist intensity on a weekday. There was

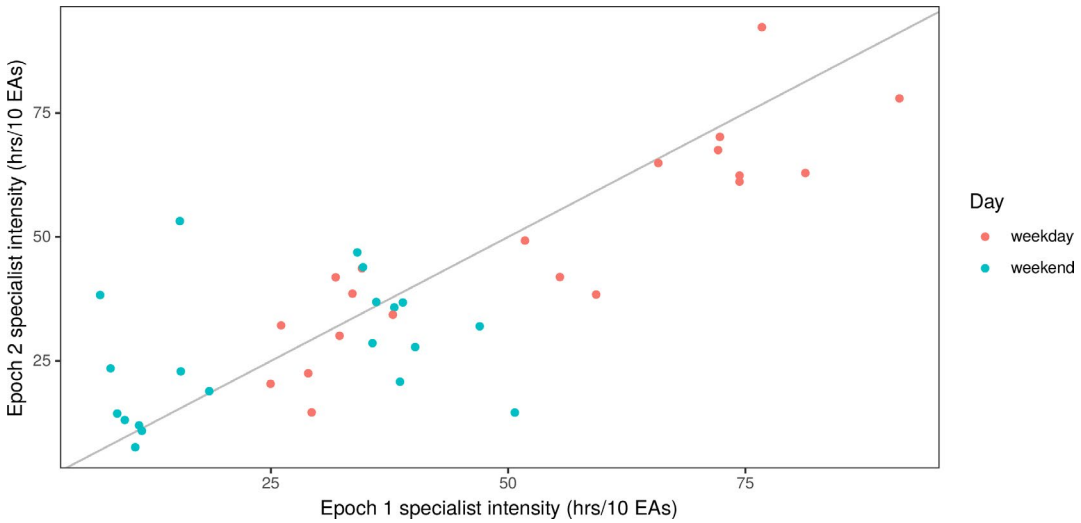

FIGURE 4 Specialist intensity before and after implementation of the 7-day services policy at weekends and on weekdays

little overall change in specialist intensity over the period of the policy and weekend levels remained consistently lower than those in the week. Across our 20 sites the mean level of specialist intensity at the weekend increased from 25.1 h per 10 emergency admissions to 27.7, but for weekdays the equivalent numbers were 52.7 and 49.8, respectively. However, there was still significant variation in specialist intensity within hospitals over time, by day of the week, and between hospitals overall.

4.2 | Model comparisons

Table 2 reports the WAIC and LOO-CV statistics for the different models. The semi-parametric baseline hazard model provides the best fit to the data. This is also reflected in Figures 5 and 6, which show the posterior predictive density for length of stay and survival function for different baseline hazards and with a linear specification for specialist intensity (all specifications of the specialist intensity function showed the same pattern). The exponential model under-predicts the hazard for short lengths of stay and the Weibull model overpredicts it. Comparison of the lower panels in each of these figures also suggests the observation of quality, as adverse events, provides little improvement in model fit for these outcomes. In terms of the polynomial specification for specialist intensity, there is little difference between linear, quadratic and cubic according to both WAIC and LOO-CV statistics. The cubic specification is generally the least preferred although the difference is relatively small. Theoretically, a quadratic or cubic model is preferred as we expect a diminishing marginal benefit of additional specialist intensity, so our selected model is one with semi-parametric baseline hazard function and a quadratic function for specialist intensity. We note that the models appear identified (in a Bayesian sense) as the posterior variance of model parameters is significantly smaller than prior variance and model parameters all appear unimodal.

4.3 | Effect of specialist intensity

Table 2 reports estimates of the direct, indirect and total effects, π_1, π_2 and π_3, respectively, associated with a change from 25 to 50 specialist hours per 10 EAs. In all models, the estimated indirect effect is very small—all estimates of π_2 are approximately one. The posterior mean (95% credible interval) change in the risk of an adverse event associated with the change in specialist intensity from the quadratic model is −0.7 (−1.5, 0.0) percentage points; the posterior mean (95% CrI) proportionate change in length of stay associated with an adverse event from the same model is 1.73 (1.40, 2.15). Both values are plausible but combine to produce a net effect on expected length of stay of less than 1% of the mean length of stay.

Estimates of the direct effect are broadly consistent between models (Table 2). Larger reductions in length of stay were estimated from models with quadratic or cubic specifications compared to a linear specification. Figure 7 shows estimates of changes to length of stay relative to 25 h per 10 EAs by specification of the specialist intensity function. While the evidence is consistent with no effect of specialist intensity, given the uncertainty intervals, there is a high probability that higher levels of specialist intensity are associated with shorter lengths of stay. The quadratic model also provides evidence for a diminishing marginal benefit of specialist intensity.

For a mean length of stay of 4.7 days and a mean cost per specialist hour of £77 (so a cost of £1925 to increase from 25 to 50 h per 10 EAs), a value of $\pi_3 = 0.91$ or lower is required for a negative expected net cost of the policy of increasing specialist intensity from 25 to 50 h per 10 EAs. Under a decision to invest if the expected net cost is negative, then the linear model would provide evidence

TABLE 2 Main results including estimated net cost of increasing specialist intensity from 25 to 50 h per 10 EAs. Reported effects and net costs are posterior means (95% credible interval)

Model	Sample	Quality	Hazard	Function	WAIC	LOO	Direct effect (π_1)	Indirect effect (π_2)	Total effect (π_3)	Net cost (£)
1	Ambulance	Y	Exponential	Linear	11,336	11,337	0.93 (0.81, 1.07)	0.99 (0.98, 1.00)	0.93 (0.81, 1.07)	66 (−146, 308)
2	Ambulance	Y	Weibull	Linear	11,340	11,341	0.93 (0.82, 1.06)	0.99 (0.98, 1.00)	0.93 (0.81, 1.06)	63 (−140, 294)
3	Ambulance	Y	Semi-parametric	Linear	9860	9861	0.92 (0.81, 1.05)	0.99 (0.98, 1.00)	0.91 (0.80, 1.04)	41 (−155, 270)
4	Ambulance	Y	Exponential	Quadratic	11,336	11,337	0.84 (0.71, 0.98)	0.99 (0.99, 1.00)	0.84 (0.71, 0.98)	−91 (−320, 156)
5	Ambulance	Y	Weibull	Quadratic	11,339	11,340	0.85 (0.71, 1.00)	1.00 (0.99, 1.00)	0.85 (0.71, 1.00)	−77 (−318, 185)
6	Ambulance	Y	Semi-parametric	Quadratic	9862	9863	0.85 (0.73, 1.00)	1.00 (0.99, 1.00)	0.85 (0.73, 1.00)	−68 (−289, 190)
7	Ambulance	Y	Exponential	Cubic	11,341	11,343	0.85 (0.71, 1.03)	1.00 (0.99, 1.00)	0.85 (0.71, 1.03)	−66 (−317, 251)
8	Ambulance	Y	Weibull	Cubic	11,344	11,347	0.85 (0.71, 1.02)	1.00 (0.99, 1.00)	0.85 (0.71, 1.02)	−70 (−318, 224)
9	Ambulance	Y	Semi-parametric	Cubic	9868	9869	0.85 (0.71, 1.02)	1.00 (0.99, 1.00)	0.85 (0.70, 1.02)	−67 (−329, 219)
10	Ambulance	N	Semi-parametric	Linear	9532	9533	NA	NA	0.92 (0.80, 1.05)	16 (−160, 289)
11	Ambulance	N	Semi-parametric	Quadratic	9532	9532	NA	NA	0.85 (0.73, 1.00)	−66 (−288, 189)
12	Ambulance	N	Semi-parametric	Cubic	9537	9538	NA	NA	0.85 (0.71, 1.02)	−72 (−324, 226)
13	All admissions	Y	Semi-parametric	Linear	16,904	16,904	0.94 (0.86, 1.03)	1.00 (0.99, 1.00)	0.94 (0.86, 1.03)	87 (−51, 237)
14	All admissions	Y	Semi-parametric	Quadratic	16,899	16,990	0.86 (0.77, 0.95)	1.00 (1.00, 1.00)	0.86 (0.77, 0.95)	−60 (−210, 96)
15	All admissions	Y	Semi-parametric	Cubic	16,902	16,903	0.82 (0.73, 0.92)	1.00 (1.00, 1.00)	0.82 (0.73, 0.92)	−122 (−281, 51)

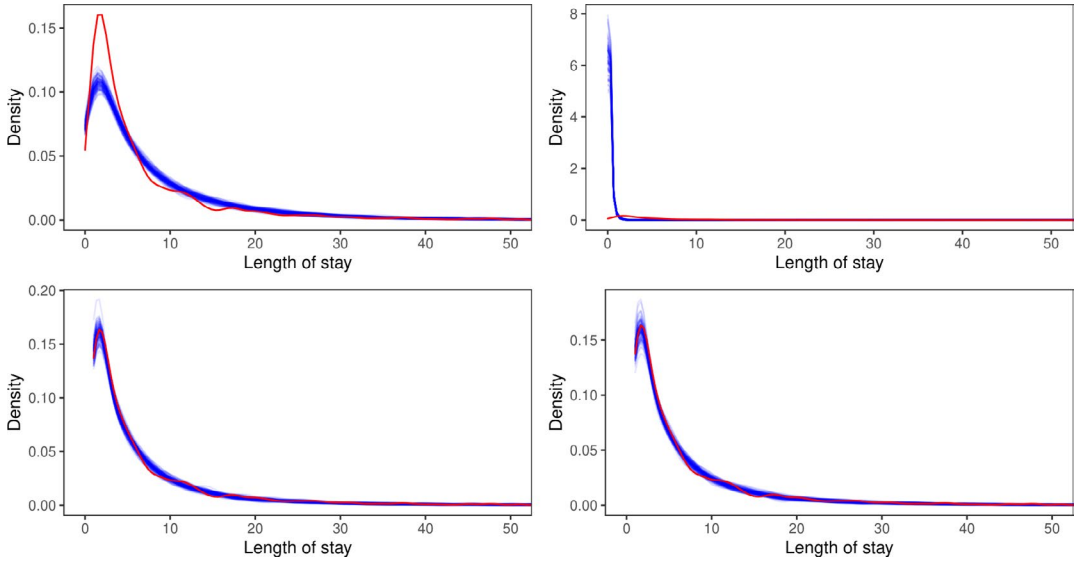

FIGURE 5 Samples from posterior predictive density of length of stay (blue lines) and empirical density of observed data (red line) for surviving patients arriving by ambulance with linear specification for specialist intensity and with exponential (top left), Weibull (top right), semi-parametric (bottom left and right) baseline hazard functions. Bottom right plot is from model without observed quality

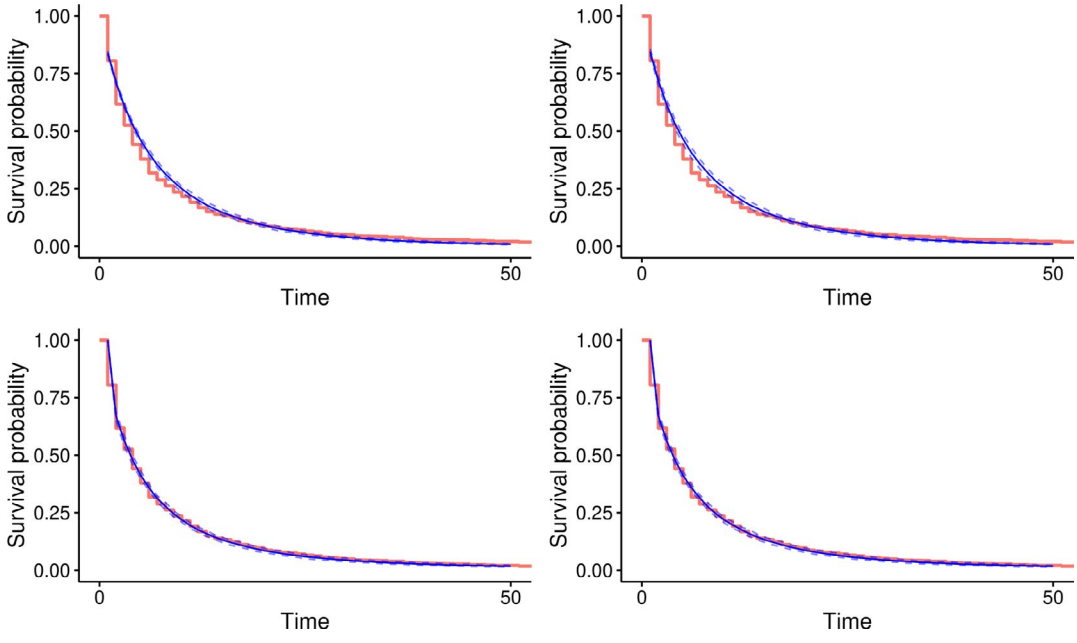

FIGURE 6 Samples from posterior predictive density of survival function (blue lines) and Kaplan-Meier estimate (red line) for patients arriving by ambulance with linear specification for specialist intensity and with exponential (top left), Weibull (top right), semi-parametric (bottom left and right) baseline hazard functions. Bottom right plot is from model without observed quality

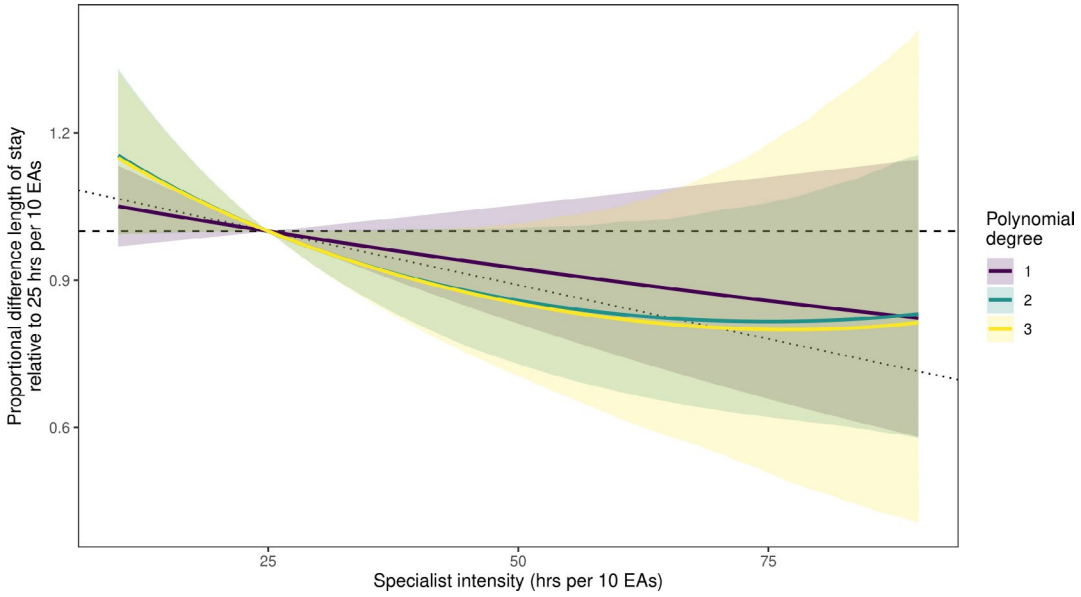

FIGURE 7 Posterior mean and 95% credible intervals of proportionate change in length of stay relative to a specialist intensity of 25 h per EA by specification of specialist intensity function for semi-parametric baseline hazard model. Dashed line indicates no difference; dotted line indicates minimum effect required for negative expected net cost

against the policy although the opposite conclusion would be reached from the quadratic and cubic models. Indeed as Figure 7 shows, the linear model suggests no change in specialist intensity would be cost saving, whereas changes up to approximately 75 h per 10 EAs are cost saving according to the other models. The probability that increasing specialist intensity from 25 to 50 h per 10 EAs is cost saving is approximately 35% for the linear model and 70% for the quadratic and cubic models.

5 | DISCUSSION

There are well-developed guidelines for the evaluation of clinical interventions, which has permitted the generation of reliable evidence to support health service treatment and investment decisions. Service delivery interventions have not generally received as much scrutiny and they present a number of distinctive challenges. There has been recent recognition of the need to develop appropriate methodology for their evaluation (Raine et al., 2016). In this article, we have considered some of the key statistical challenges relating to the assessment of an intervention's effects on patient length of stay and we have proposed an appropriate empirical strategy and model from which we draw a number of conclusions.

Many analyses of hospital-level interventions rely on data from admitted patients. Selection bias will result if interventions affect which patients are admitted to the hospital. The bias can be addressed if there are data on all hospital attendances, and prospective studies of service delivery interventions should aim to collect this information. However, these data may not be available to researchers. We have proposed an approach for when only data from admissions are available. If the proportional hazards assumption holds, which can be examined graphically, then relative treatment effects estimated from one subsample will apply to the whole sample. Changes to the probability of admission are likely

to affect those at the margin so the admission of severely ill patient groups, such as those arriving by ambulance or those requiring intensive care, is likely to be independent of any intervention. Our approach also provides a simple means of decomposing direct and indirect effects, which also relies on the proportional hazards assumption. In the absence of this assumption and using an alternative model, it would still be possible to draw from the posterior predictive distribution of length of stay for both direct and indirect effects (Equation 1) to enable assessment of treatment effects. However, these would not be generalisable to other subsamples and so would limit their use. Further research is required into methods of evaluation for service delivery intervention for non-proportional hazards.

Our analysis suggests that any change in patient length of stay associated with specialist intensity is almost exclusively attributable to a direct, as opposed to indirect, effect. This is likely due to earlier patient discharges as only specialists are able to discharge patients. Our evidence is consistent with previous studies of the 'weekend effect' that have suggested that any change in patient health status due to day of admission is likely to be small (Meacock & Sutton, 2018; Meacock et al., 2017; Mohammed et al., 2017). Indeed, preventable harm is relatively rare among patients; Hogan et al. (2012) estimates only around 5% of in-hospital deaths are preventable, for example. Any intervention that targets quality would therefore need to have a very large relative effect to be both distinguishable from noise in a statistical evaluation and relevant to policy makers. Our models provide evidence that greater specialist intensity is associated with reduced risk of adverse events, and adverse events increase length of stay, but the net indirect effect is relatively very small. As a result, there was little qualitative difference in the results from models where it was and was not observed. However, this does not imply that observations of quality, such as adverse events or errors in care, are not required for the reliable evaluation of any service delivery intervention.

We examined both parametric and semi-parametric hazard functions. The semi-parametric models provided a much better fit to the data in this study when compared to the exponential and Weibull parametric models. There has been a growing use of semi-parametric and non-parametric specifications of hazard models in the survival analysis and competing risks literature including for Bayesian approaches, such as using a constant piecewise hazard formulation with discrete time intervals or spline-based models like the one used in our analysis (e.g., Lee et al., 2015; Song et al., 2002; Wang & Taylor, 2001; Zhang et al., 2017). The Cox proportional hazard model is an alternative and popular method of estimating model parameters without parametric specification of a baseline hazard model (Cox, 1972), for which Bayesian formulations also exist (Vallejos & Steel, 2017). However if interest lies in the marginal distribution of length of stay or its joint distribution with another outcome then a full hazard model needs to be specified and the semi-parametric shared parameter formulation used in this article provides a means of doing so.

The results are clearly sensitive to modelling assumptions. Under a linear specification of specialist intensity we would conclude that the intervention is not cost saving, whereas it would be under a higher-order polynomial specification. We have opted to use a combination of Bayesian goodness of fit statistics, graphical posterior predictive checks, and theoretical considerations to select between candidate models, which would lead us to conclude that the intervention would be cost saving. Alternatively, model-averaging methods could be used. Jackson et al. (2010) describe a method to resolve structural uncertainty in cost-effectiveness models using a Bayesian model-averaging approach, weighting results from models by their relative probabilities given the data. Similar methods such as the 'stacking' of predictive distributions could be used to combine estimates (Yao et al., 2018).

The conclusions here are only as good as the data that drive them. Adverse events were determined by a subjective case note review, the methodology of which followed standard practice for studies of patient safety (Hogan et al., 2015). In total, 800 records were reviewed twice by different reviewers, and reliability and agreement between reviewers was generally assessed to be low, reflecting results

from similar studies (e.g. Hogan et al., 2015). Reviewers were blinded to specialist intensity and day of admission, so any errors were likely independent of the other variables informing our models. One might assume that the adverse events identified by clinical reviewers were no better than random guessing and contained no reliable information on preventable harm. However, the estimated effects of specialist intensity on the risk of adverse events, and of adverse events on length of stay, had high face validity both in terms of direction and magnitude. We believe that adverse events in this analysis can therefore be interpreted as a noisy measure of hospital quality and they do not introduce any systematic bias. However, the lack of agreement could be due broader misclassification errors, which could affect both the 'sensitivity' (correctly identifying patient harm) and 'specificity' (correctly identify cases where no harm occurred) of the review process. Depending on the nature of these errors they could introduce large biases (Greenland, 1996), which represents a potentially significant weakness of this study, and indeed any based on case note review. However, little research on the nature and consequences of misclassification error in case note review exists since there is no 'Gold standard' against which to compare the results. Future research is needed to resolve these issues.

Response rates were 40–50% for the annual survey of specialist hours (Aldridge et al., 2016). It is possible that specialists who practice more hours may be more likely to respond, or vice versa. But we do not see any strong reason to suspect that there would be any correlation between overall specialist presence in a hospital and the propensity to respond to the survey. However, caution should be exercised in the interpretation of the results. There also was a lack of information on the distribution of clinician salaries so we did not model the uncertainty in costs. From a decision theoretic perspective, it is the expected net cost that matters for the decision, which would be unaffected by adding uncertainty to the costs if the mean were left unchanged. However, it may have some effect on the probability estimate, although we do not believe these effects to be large.

One of the key aims of the 7-day services policy was to increase the level of weekend specialist intensity to the weekday level. The policy was not implemented as intended, however, and specialist intensity remained lower at the weekend in the post-policy period. The evidence from this article suggests that had levels increased as intended then it would have likely been cost saving. It would cost approximately £1.9 million in new expenditure to increase specialist intensity by 25 h per 10 EAs at the weekend in a hospital with 10,000 weekend admissions a year (assuming a mean cost per specialist hour of £77), but the expected net cost to the hospital would be £-1.1 million on the results from our preferred model (Table 2). However, our cost analysis is of the effects of increasing specialist intensity for a *fixed* patient population. Should the length of stay of patients be reduced, this may lead to a change in patient admissions, throughput, or referral patterns to use the extra capacity. From an economic perspective, this is the problem of supply-induced demand. Our model does not allow for this type of feedback given that it is based on acyclic graphs and the question remains as to the extent of this effect, however we expect it to be fairly minimal relative to the scale of change of the intervention. Furthermore, given the uncertainty around net costs a full cost-benefit analysis may be warranted, including assessment of the health consequences of adverse events (Watson et al., 2018).

ACKNOWLEDGEMENTS

The authors thank the HiSLAC team for their work in collecting, preparing, and reviewing the data in the project. The HiSLAC team excluding the authors are: Cassie Aldridge (University of Birmingham), Alan Girling (University of Birmingham), Gavin Rudge (University of Birmingham), Carolyn Tarrant (University of Leicester), Liz Sutton (University of Leicester) Janet Willars (University of Leicester), Chris Beet (Queen Elizabeth Hospital Birmingham, UK), Amunpreet Boyal (Queen Elizabeth Hospital Birmingham, UK), Peter Rees (Academy of Medical Royal Colleges, UK), Chris Roseveare (Southern Health NHS Foundation Trust), Mark Temple (Queen Elizabeth Hospital Birmingham, UK) Yen-Fu

Chen (University of Warwick), Michael Clancy (United Hospitals Southampton NHSFT), Louise Rowan (University of Birmingham), Joanne Lord (University of Southampton, UK), Russell Mannion (University of Birmingham), Timothy Hofer (University of Michigan Division of General Medicine).

Local project leads in each of the 20 hospitals: Professor Mark O'Donnell, Dr Chris Adcock, Dr Paul Peter, Dr Nnenna Osuji, Dr Emma Rowland, Dr William Bernal, Dr Mehool Patel, Dr Jayachandran Radhakrishnan, Dr Emma Vaux, Dr Rupert Negus, Dr Stuart Henderson, Dr Andrew Gibson, Dr Richard Heinink, Dr Hassan Paraiso, Dr Earl Williams, Dr Lee Dowson, Dr Sanjiv Jain, Dr Mike Berry, Dr Catherine Snelson, Dr Becky Thorpe, Dr Emma Redfern, Dr Emma Rowlandson. Prof Nick Black and Dr Helen Hogan for access to their case record review methodology. Sue Pargeter, Manager, HS&DR Programme.

CONFLICT OF INTEREST

None declared.

REFERENCES

Aldridge, C., Bion, J., Boyal, A., Chen, Y.F., Clancy, M., Evans, T. et al. (2016) Weekend specialist intensity and admission mortality in acute hospital trusts in England: a cross-sectional study. *The Lancet*, 388, P178–186.

Anselmi, L., Meacock, R., Kristensen, S.R., Doran, T. & Sutton, M. (2017) Arrival by ambulance explains variation in mortality by time of admission: retrospective study of admissions to hospital following emergency department attendance in England. *BMJ Quality & Safety*, 26, 613–621. http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2016-005680.

Bell, C.M. & Redelmeier, D.A. (2001) Mortality among patients admitted to hospitals on weekends as compared with weekdays. *New England Journal of Medicine*, 345, 663–668. http://www.nejm.org/doi/abs/10.1056/NEJMsa003376.

Bion, J., Aldridge, C.P., Girling, A., Rudge, G., Beet, C., Evans, T. et al. (2017) Two-epoch cross-sectional case record review protocol comparing quality of care of hospital emergency admissions at weekends versus weekdays. *BMJ Open*, 7, e018747.

Brennan, T.A., Leape, L.L., Laird, N.M., Hebert, L., Localio, A.R., Lawthers, A.G. et al. (1991) Incidence of adverse events and negligence in hospitalized patients. *New England Journal of Medicine*, 324, 370–376. http://www.nejm.org/doi/abs/10.1056/NEJM199102073240604.

Brilleman, S., Elci, E., Novic, J. & Wolfe, R. (2020) Bayesian survival analysis using the rstanarm R package. https://arxiv.org/abs/2002.09633.

British Medical Association. (2019) Pay scales for consultants in England. https://www.bma.org.uk/advice/employment/pay/consultants-pay-england.

Chen, Y.-F., Armoiry, X., Higenbottam, C., Cowley, N., Basra, R., Watson, S.I. et al. (2019) Magnitude and modifiers of the weekend effect in hospital admissions: a systematic review and meta-analysis. *BMJ Open*, 9, e025764. http://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2018-025764.

Cook, R.J. & Lawless, J.F. (1997) Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16, 911–924. https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19970430)16:8%3C911::AID-SIM544%3E3.0.CO;2-I.

Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–202. http://doi.wiley.com/10.1111/j.2517-6161.1972.tb00899.x.

Department of Health. (2015) 7-day NHS services: a factsheet. Technical report, Department of Health & Social Care, London, UK. https://www.gov.uk/government/publications/7-daynhs-%0Aservices-a-factsheet/7-day-nhs-services-a-factsheet.

Freemantle, N., Richardson, M., Wood, J., Ray, D., Khosla, S., Shahian, D. et al. (2012) Weekend hospitalization and additional risk of death: an analysis of inpatient data. *Journal of the Royal Society of Medicine*, 105, 74–84.

Freemantle, N., Ray, D., McNulty, D., Rosser, D., Bennett, S., Keogh, B.E. et al. (2015) Increased mortality associated with weekend hospital admission: a case for expanded seven day services? *BMJ*, 351, h4596.

Fulcher, I.R., Shpitser, I., Marealle, S. & Tchetgen Tchetgen, E.J. (2020) Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 199–214.

Galloway, M., Hegarty, A., McGill, S., Arulkumaran, N., Brett, S.J. & Harrison, D. (2018) The effect of ICU out-of-hours admission on mortality. *Critical Care Medicine*, 46, 290–299. http://insights.ovid.com/crossref?an=00003246-201802000-00015.

Gaynor, M., Propper, C. & Seiler, S. (2016) Free to choose? Reform, choice, and consideration sets in the English National Health Service. *American Economic Review*, 106, 3521–3557. https://pubs.aeaweb.org/doi/10.1257/aer.20121532.

Gelman, A., Jakulin, A., Pittau, M.G. & Su, Y.S. (2008) A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2, 1360–1383.

Girling, A.J., Hofer, T.P., Wu, J., Chilton, P.J., Nicholl, J.P., Mohammed, M.A. et al. (2012) Case-mix adjusted hospital mortality is a poor proxy for preventable mortality: a modelling study. *BMJ Quality & Safety*, 21, 1052–1056. http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2012-001202.

Gowrisankaran, G. & Town, R.J. (1999) Estimating the quality of care in hospitals using instrumental variables. *Journal of Health Economics*, 18, 747–767. https://linkinghub.elsevier.com/retrieve/pii/S0167629699000223.

Graham, P.L., Ryan, L.M. & Luszcz, M.A. (2011) Joint modelling of survival and cognitive decline in the Australian Longitudinal Study of Ageing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60, 221–238. http://doi.wiley.com/10.1111/j.1467-9876.2010.00737.x.

Greenland, S. (1996) Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25, 1107–1116. https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/25.6.1107-a.

Henderson, R., Diggle, P.J. & Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1, 465–480.

Hess, K.R. (1995) Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine*, 14, 1707–1723.

Hogan, H., Healey, F., Neale, G., Thomson, R., Vincent, C. & Black, N. (2012) Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Quality & Safety*, 21, 737–745. http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2011-001159.

Hogan, H., Zipfel, R., Neuburger, J., Hutchings, A., Darzi, A. & Black, N. (2015) Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. *BMJ*, 351, h3239. http://www.bmj.com/lookup/doi/10.1136/bmj.h3239.

Jackson, C.H., Sharples, L.D. & Thompson, S.G. (2010) Structural and parameter uncertainty in Bayesian cost-effectiveness models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59, 233–253. http://doi.wiley.com/10.1111/j.1467-9876.2009.00684.x.

Jena, A.B., Prasad, V., Goldman, D.P. & Romley, J. (2015) Mortality and treatment patterns among patients hospitalized with acute cardiovascular conditions during dates of National Cardiology Meetings. *JAMA Internal Medicine*, 175, 237. http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2014.6781.

Kerlin, M.P., Small, D.S., Cooney, E., Fuchs, B.D., Bellini, L.M., Mikkelsen, M.E., et al. (2013) A randomized trial of nighttime physician staffing in an intensive care unit. *New England Journal of Medicine*, 368, 2201–2209. http://www.nejm.org/doi/10.1056/NEJMoa1302854.

Lee, K.H., Haneuse, S., Schrag, D. & Dominici, F. (2015) Bayesian semiparametric analysis of semicompeting risks data: investigating hospital readmission after a pancreatic cancer diagnosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64, 253–273. http://doi.wiley.com/10.1111/rssc.12078.

Lilford, R.J., Chilton, P.J., Hemming, K., Girling, A.J., Taylor, C.A. & Barach, P. (2010) Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. *BMJ*, 341, c4413. http://www.bmj.com/cgi/doi/10.1136/bmj.c4413.

Long, J.D. & Mills, J.A. (2018) Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Medical Research Methodology*, 18, 138. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0592-9.

Meacock, R. & Sutton, M. (2018) Elevated mortality among weekend hospital admissions is not associated with adoption of seven day clinical standards. *Emergency Medicine Journal*, 35, 108–113.

Meacock, R., Anselmi, L., Kristensen, S.R., Doran, T. & Sutton, M. (2017) Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission. *Journal of Health Services Research and Policy*, 22, 12–19.

Meacock, R., Anselmi, L., Kristensen, S. R., Doran, T. & Sutton, M. (2019) Do variations in hospital admission rates bias comparisons of standardized hospital mortality rates? A populationbased cohort study. *Social Science & Medicine*, 235, 112409. https://linkinghub.elsevier.com/retrieve/pii/S0277953619303958.

Mohammed, M.A., Deeks, J.J., Girling, A., Rudge, G., Carmalt, M., Stevens, A.J. et al. (2009) Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. *BMJ*, 338, 780. http://www.bmj.com/cgi/doi/10.1136/bmj.b780.

Mohammed, M., Faisal, M., Richardson, D., Howes, R., Beatson, K., Speed, K. et al. (2017) Impact of the level of sickness on higher mortality in emergency medical admissions to hospital at weekends. *Journal of Health Services Research and Policy*, 22, 236–242.

Montgomery, J.M., Nyhan, B. & Torres, M. (2018) How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62, 760–775. http://doi.wiley.com/10.1111/ajps.12357.

NHS England. (2013) NHS services, seven days a week forum: summary of initial findings. Techical report, London, UK.

NHS England. (2019) NHS reference costs 2017/18. https://improvement.nhs.uk/resources/reference-costs/#rc1718.

Pearl, J. (2012) The causal foundations of structural equation modeling. In: Hoyle, R. (Ed.) *Handbook of structural equation modeling*. New York: Guilford Press, pp. 68–91.

Raine, R., Fitzpatrick, R., Barratt, H., Bevan, G., Black, N., Boaden, R. et al. (2016) Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Services and Delivery Research*, 4, 1–136. https://www.journalslibrary.nihr.ac.uk/hsdr/hsdr04160/.

Rosenbaum, P.R. (1984) The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147, 656. https://www.jstor.org/stable/10.2307/2981697?origin=crossref.

Schoenfeld, D. (1982) Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239. https://www.jstor.org/stable/2335876?origin=crossref.

Song, X., Davidian, M. & Tsiatis, A.A. (2002) A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58, 742–753. http://doi.wiley.com/10.1111/j.0006-341X.2002.00742.x.

Sun, J., Girling, A.J., Aldridge, C., Evison, F., Beet, C., Boyal, A. et al. (2019) Sicker patients account for the weekend mortality effect among adult emergency admissions to a large hospital trust. *BMJ Quality & Safety*, 28, 223–230. http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2018-008219.

Vallejos, C.A. & Steel, M.F.J. (2017) Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 613–631. http://doi.wiley.com/10.1111/rssa.12211.

Vehtari, A., Gelman, A. & Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413–1432.

Wang, Y. & Taylor, J.M.G. (2001) Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96, 895–905. http://www.tandfonline.com/doi/abs/10.1198/016214501753208591.

Watson, S., Arulampalam, W. & Petrou, S. (2017) The effect of health care expenditure on patient outcomes: evidence from English neonatal care. *Health Economics*, 26, e274–e284. http://doi.wiley.com/10.1002/hec.3503.

Watson, S.I., Chen, Y.-F., Bion, J.F., Aldridge, C.P., Girling, A. & Lilford, R.J. (2018) Protocol for the health economic evaluation of increasing the weekend specialist to patient ratio in hospitals in England. *BMJ Open*, 8, e015561. http://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2016-015561.

Wilcox, M.E., Chong, C.A.K.Y., Niven, D.J., Rubenfeld, G.D., Rowan, K.M., Wunsch, H. et al. (2013) Do intensivist staffing patterns influence hospital mortality Following ICU admission? A systematic review and meta-analyses*. *Critical Care Medicine*, 41, 2253–2274. https://insights.ovid.com/crossref?an=00003246-201310000-00001.

Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. (2018) Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13, 917–1007. https://projecteuclid.org/euclid.ba/1516093227.

Zeng, D. & Lin, D.Y. (2009) Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics*, 65, 746–752. http://doi.wiley.com/10.1111/j.1541-0420.2008.01126.x.

Zhang, S., Müller, P. & Do, K.-A. (2010) A Bayesian semiparametric survival model with longitudinal markers. *Biometrics*, 66, 435–443. http://doi.wiley.com/10.1111/j.1541-0420.2009.01276.x.

Zhang, D., Chen, M.-H., Ibrahim, J.G., Boye, M.E. & Shen, W. (2017) Bayesian model assessment in joint modeling of longitudinal and survival data with applications to cancer clinical trials. *Journal of Computational and Graphical Statistics*, 26, 121–133. https://www.tandfonline.com/doi/full/10.1080/10618600.2015.1117472.

How to cite this article: Watson SI, Lilford RJ, Sun J, Bion J. Estimating the effect of health service delivery interventions on patient length of stay: A Bayesian survival analysis approach. *J R Stat Soc Series C*. 2021;00:1–23. https://doi.org/10.1111/rssc.12501