# Cover page

Intentionally left blank for now

# Title page

## Working title

Delay to admission to critical care, unit occupancy, and mortality for deteriorating ward patients: a prospective observational cohort study

## Authors

* Steve Harris
* Mervyn Singer
* David Harrison
* Kathy Rowan
* Colin Sanderson

## Acknowledgements

* Sarah Power
* Sheila Harvey
* Rahi and Emma
* Participating hospitals

# Introduction

Recent National Health Service (NHS) policy highlights the importance of identifying and responding to the deteriorating ward patient.[2007, #773] Current guidelines recommend that critical care admission should be delivered within four hours[2013, #10047; Dellinger et al., 2013, #14540], but unbiased evaluations of this are difficult. Because benefit is presumed, randomised evaluations of early admission to critical care are unethical, yet without quantification it is difficult to assess the magnitude and importance of this problem.

Non-randomised evaluations will be primarily confounded by treatment allocation bias. Patients are prioritised based on clinical severity so early admissions would be anticipated to be more unwell with worse outcomes. Risk adjustment will help but depends on accurately measuring all factors that lead to the treatment allocation decision. It is likely that this measured severity is an incomplete description, and that there are other ‘end of the bed’ factors that prompt clinicians to recommend critical care.

An alternative to risk adjustment is to seek an instrument that naturally randomises patients. We, like others[Shmueli et al., 2004, #16778], have used critical care occupancy for this purpose. High critical care occupancy is likely to mean that some patients are turned away, and some are admitted with delay.[Stelfox et al., 2012, #4003] We assume that occupancy can *only* cause harm by deflecting or delaying admissions to exclude confounding. We can then compare the outcomes of those patients who happen to deteriorate when a critical care bed is available to those who are unlucky, and deteriorate when the critical care unit is full. Any difference in patient survival can be ascribed to the the effect of deflecting or delaying the admission. To get at the specific effect of delayed admission, we repeat the analysis amongst those patients offered a critical care bed, wherein only a small proportion of the controls will not be admitted.

This approach, known as Instrumental Variable (IV) analysis, has been previously used to remove unmeasured confounding in the assessment of `flu vaccine efficacy and cardiac catheterisation.[Stukel et al., 2007, #16499; Wong et al., 2012, #16498] The IV analysis simply replaces the strong assumption of no unmeasured confounding, upon which risk adjustment rests, with, we believe, the more reasonable assumption that critical care occupancy has no other direct effect on the outcome of a deteriorating ward patient.[Hernan and Robins, 2006, #12386] The additional benefit of this approach is that the effect of critical care occupancy on the admission pathway for deteriorating ward patients is of direct interest itself, with an obvious implication for health care provision.

Where critical care provision is more generous, this question would be more difficult to answer because delay arising from a lack of critical care capacity would be rare. However, the United Kingdom (UK) is ranked 24th out of 31 European countries in terms of provision of critical care per capita population, and compares similarly unfavourably with North American health care systems.[Rhodes et al., 2012, #15692; Wunsch et al., 2008, #761] This presents therefore an opportunity to measure the effect of early admission to critical care.

# Methods

The study protocol is available [here](https://www.icnarc.org/Our-Research/Studies/Spotlight/About).

## Study design, participants and procedures

The (SPOT)light study was a prospective observational cohort study of the deteriorating ward patient referred to critical care. Patients were eligible if they were inpatients on general hospital wards who had been referred to, and assessed at the bedside by, a member of the critical care staff, or the critical care outreach team (CCOT). Repeat visits, re-admissions, cardiac arrests, and deaths during the visit were excluded as were admissions following surgery (where delay was due to the process of care), and patients with pre-existing treatment limitations. One hospital subsequently found to have critical care capacity in addition to that monitored by the CMP was excluded.

The study was registered on the National Institute of Health Research (NIHR) research portfolio, and hospitals were eligible if they participated in the Intensive Care National Audit & Research Centre’s Case Mix Programme Database (ICNARC CMPD). Research teams at each hospital attended a Dataset Familiarisation Course, and a data collection manual containing definitions of items to be collected was provided. The Clinical Trials Unit at ICNARC provided support for research queries during the study.

Hospitals were asked to report all consecutive ward referrals. Contemporaneous data collection was recommended, but missed referrals were sought and accepted retrospectively. We used the proportion of emergency ward admissions in the ICNARC CMP that were successfully linked to the (SPOT)light database to quality control the study each month. All data from individual months where data linkage fell below a threshold of 80% were excluded. Reporting was via a secure on-line web portal which performed real-time field and record level validation. Further on-line validation reports were completed by all hospitals before the database was locked in September 2012. Fact and date of death were then requested from the NHS Information Service. CCOT provision was reported by participating hospitals, and CMP data and Hospital Episode Statistics (HES) were used to define critical care provision, occupancy, and hospital characteristics.

## Definitions

Physiology measurements at the time of the ward assessment were abstracted, and the ICNARC physiology score, NHS National Early Warning Score, and the SOFA score were calculated assigning missing values zero weights as recommended.[Harrison et al., 2007, #1640; Physicians, 2012, #9726; Vincent et al., 1996, #719] Existing dependency was defined using the Critical Care Minimum Dataset (CCMDS) levels of care: levels 0 and 1 are most commonly provided on normal wards; and, levels 2 and 3 within high dependency (HDU) and intensive care units (ICU) respectively.[2009, #18015]

Critical care unit occupancy was calculated as the difference between the maximum number of beds reported to the ICNARC CMP, and the number of actively treated patients occupying those beds in the same hour that the ward patient was assessed. Units were defined as having two or more, one, or zero or fewer empty beds (if occupancy exceeded reported capacity).

At the bedside assessment we recorded the level of care recommended, and the decision to admit. We defined early admission as one occurring within four hours of the bedside assessment in line with recently published UK guidelines.[2013, #10047]

## Statistical analysis

The primary analysis evaluated the effect of critical care admission within four hours of the bedside assessment on 90 day mortality. The controls included all patients assessed and not admitted early. This effect measure summarises both the effect of delayed admission, and the effect of never being admitted. Assuming that most patients offered critical care would be admitted, we performed a secondary analysis within this subgroup. This was intended to isolate the specific contrast between early and delayed admission (without contamination from those never admitted).

We first built proportional hazards models, and used risk adjustment to handle the anticipated treatment selection bias before resorting to the IV analysis. These latter models are constructed in two stages. In the first stage, a selection model is constructed to predict early admission including the effect of occupancy. In the second stage, an outcome model is built replacing occupancy with fitted prediction of early admission from the selection model. All models are adjusted for patient level confounders including age, sex, reported referral delay, sepsis diagnosis, peri-arrest status, current and recommended CCMDS level of care, and severity of illness using the ICNARC physiology score.[Harrison 2007]

In order to estimate the total effect of early admission, the IV selection model excludes factors that act through the primary exposure (e.g. if small critical care units are more often full then including this in the selection model would mean that we would only retrieve the effect of occupancy after allowing for unit size). However, we were also interested in how these unit and hospital characteristics affected the recommendation, the decision to admit, and the early delivery of critical care. These intermediate outcomes therefore became the dependent variables in three logistic regression models that included patient level, hospital level (teaching status, number of overnight admissions, emergency case-mix, CCOT provision, ward to critical care referral rate), unit level (beds and turnover — as admissions per bed per month), and timing factors (out-of-hours from 7pm-7am), weekend, and winter periods.

We tested for weak instruments using the the Kleibergen-Paap F test [Kleibergen and Paap, 2006, #17245], and used Huber-White (robust) standard errors to allow for hospital-level clustering and potential heteroscedasticity. Bivariate probit IV models were used to ensure that model predictions are correctly constrained. These report on the probit scale which represents a change in the Z-score of the distribution of the prediction. To aid interpretation, we also calculated the marginalised average treatment effect (ATE) for the population, and converted coefficients to approximate odds ratios (OR) by scaling by 1.6.[Amemiya, 1981, #36641]

## Implementation

The study is registered with [ClinicalTrials.gov](http://clinicaltrials.gov/show/NCT01099813), number NCT01099813. Sample size was calculated to evaluate mortality increases from delay to admission using estimates from 2007 ICNARC CMP data. The target sample size was 12,075–20,125 patients referred to critical care allowing for delays to occur in 10–40% of admissions and mortality effect sizes of 5–10%.

Categorical data were reported as counts and percentages, and continuous data as mean (SD) or median (IQR) values. Effect measures are reported with their 95% confidence intervals. Analyses were performed in R (Version 3.03) except for the IV analysis which used the ivregress and biprobit commands provided in Stata (Version 12.1).

## Role of the funding source

The study was centrally funded by the Wellcome Trust, sponsored by ICNARC, and supported at NHS hospitals through the National Institute of Health Research service support costs. The funders of the study had no role in the study design; gathering, analysis, and interpretation of the data; writing of the report; and decision to submit the report for publication. The corresponding author had full access to all the data (including statistical reports and tables); takes responsibility for the integrity of the data and the accuracy of the data analysis; and takes final responsibility for the decision to submit for publication.

# Results

Between September 2010 and December 2011, 49 hospitals reported 21137 consecutive bedside assessments of deteriorating ward patients by the critical care team. Of these, 6038 assessments did not meet the inclusion criteria, of which the majority were follow-up assessments (1878, 31%), or patients with treatment limitation orders (2933, 49%). We used the national audit of critical care (ICNARC CMP) to quality control these data by cross checking audited admissions against reported assessments each month. We therefore excluded 66 (15%) of the 446 months we observed (where less than 80% of eligible admissions were reported). One site was excluded entirely because unreported critical care admissions to beds not monitored by the ICNARC CMP were possible. Mortality or timing data were unavailable for a further 402 patients leaving 12495 patients in 48 hospitals for analysis (Figure 1). Hospitals contributed a median of 224 patients (IQR 144–304) over 8 months (IQR 5–9), and captured 93% of eligible admissions (IQR 89–97%).

10 hospitals were university-affiliated. Critical care outreach was available in all but two hospitals. In the remainder, it was available weekdays only in 13 hospitals, seven days in 19 hospitals, and 24 hours and seven days per week in 14 hospitals. There were a median of 12 (IQR 9–18) critical care beds (mixed Level 2 (typically intensive monitoring or single organ support) and Level 3 (ventilated or multiple organ support), most often in a single physical location (45 hospitals). These units admitted 20 emergency ward admissions per month (IQR 14–26) which represented about a third (36%) of all admissions (IQR 31–43%).

Table 1 shows the baseline data for all ward patients assessed. Sepsis was reported in (3822, 31%) patients; the respiratory system was considered the source in half (3887, 51%). Organ dysfunction, defined as a SOFA score greater than one, was already present in 9862 (79%) of patients. 6726 patients (54%) were in respiratory failure, 2419 (19%) were in renal failure, and 3681 (29%) were shocked. There was a clear correlation between physiological severity and outcome (supplementary Figure 1), but pre-existing organ support was unusual (711 patients, 6%).

3279 (26%) patients were offered critical care at the bedside assessment. These patients were younger (by 1.5 years [95%CI 0.7–2.2]), and more unwell (4.4 ICNARC physiology points [95%CI 4.1–4.8]). Patients older than 80 years were less likely to be offered admission (OR 0.55 [95%CI (0.47–0.65)]) even after risk adjustment (Table 2).

In the week following the bedside assessment, 91.0% of those offered critical care were admitted (2984 patients), 40 (1.2%) died before admission, and 255 (7.8%) survived without admission. In the same period, of the 9216 patients initially refused critical care, 295 (9.0%) were subsequently admitted, and 747 (8.1%) died on the ward without admission. Over half of all deaths in the first week (787 deaths, 46.8%) occured in patients not admitted to critical care.

2492 (20%) patients were admitted early (within four hours of the bedside assessment). Of those offered critical care, 2241 patients (68%) were admitted early with a median delay of 2 hours (IQR 1–3). For those initially refused, 1038 patients (32%) were nonetheless admitted within four hours (median 12 hours, IQR 5–29 hours,Figure 2A).

Critical care occupancy at the time of the bedside assessment affected both the likelihood of being admitted to critical care, and the timing of that admission (Table 3). There were 996 (8%) patients assessed when the unit was full, 1360 (11%) assessed when there was one bed available, and 10139 (81%) assessed when there were at least two beds. Comparing the first and the last categories, fewer were offered critical care (19% [185 patients] versus 28% [2794 patients]), admissions were delayed more (5·0 hours [IQR 2·0–15·0] versus 3·0 hours [IQR 1·0–8·0], Figure 2B), and patients deteriorated further while waiting (an additional 1·5 ICNARC physiology points [95%CI 0·5–2·6]). Overall, fewer patients received critical care in the week following the initial assessment (25% [253 patients] versus 38% [3856 patients]), and a greater proportion died on the ward without admission (9% [92 deaths] versus 6% [622 deaths], Figure 3). Notably, occupancy did not affect the clinician’s bedside recommendation for critical care (36% [361 patients] when no beds were available versus 38% [3828 patients] otherwise). All of these relationships were preserved after adjustment (Table 3).

3797 (30%) patients died within 90-days. Nearly half of these deaths (46%, 1763 patients) occurred in the first week giving a 7-day mortality of 14%. There was no readily identifiable subgroup without substantial mortality. For example, 90-day mortality was 19% (501 deaths) for patients without organ dysfunction, 27% (2108 deaths) for patients not recommended for critical care, and 26% (447) for patients deemed not to warrant further critical care follow-up when assessed.

We investigated the effect of early admission (within four hours) on survival. An unbiased direct comparison of early versus late admission is not possible. There are 10003 possible controls (patients not admitted early) among whom 2002 patients are admitted late (between 4 hours and 7 days). Of the remaining 8001 possible controls, 787 (10%) died and 7214 (90%) survived without critical care admission. Excluding these would create either survival bias (by removing the most unwell), or exclusion bias (by removing the least unwell) respectively. We therefore compared the survival of the 2492 patients admitted early to all 10003 possible controls. Those admitted early were more unwell (4·4 ICNARC physiology points [95%CI 4·1–4·8]), and unadjusted 90-day survival was worse (HR 1·45 [95%CI 1·35–1·57]). After risk adjustment for measured differences in severity of illness, survival was now equivalent (HR 1·00 [95%CI 0·88–1·13] p=0·988) for early admissions and controls (Supplementary Table 1).

However, we considered it unlikely that we had completely captured patient severity, and that risk adjustment was likely incomplete. Because critical care occupancy has a strong negative correlation with early admission (Kleibergen-Paap F-statistic 70), we proceeded with the instrumental variable analysis. We built a two-stage model (bivariate probit) using the same covariates as for the survival model but, in addition, used critical care occupancy in the first stage to predict early admission, and these predictions (in the second stage) to estimate the effect of early admission on 90-day mortality (Table 4). The probit coefficient (change in Z-score of the predicted 90-day mortality) for early admission was -0·32 (95%CI -0·63 to -0·01, p= 0·043). This is approximately equivalent to an odds ratio of 0·60 (95%CI 0·37–0·98), or an average reduction in 90-day mortality of 10·0% (95%CI 0·4%–19·8%) for those admitted early.

To evaluate the effect of delay alone, we repeated the above analyses in the subgroup of 3279 patients offered critical care thus using a cohort wherein controls would be predominantly delayed admissions rather than non-admissions (supplementary Figure 2). Although (91% [2986 patients]) of this subgroup were admitted, a greater number of non-admissions were still found in the controls (295, 28%), of which 40 were deaths before critical care admission. Estimates from the survival model (HR 0·98 [95%CI 0·87-1·11], Supplementary Table 1), and the bivariate probit IV model (probit coefficient -0·17 [95%CI -0·34 to 0·01, p= 0·063], Table 4) were broadly similar. The approximate odds ratio derived from the bivariate probit model was now 0·45 (95%CI 0·19–1·07), and average mortality reduction for those admitted early was 16·7% (95%CI -0·9% to 34·0%).

# Discussion

This prospective study describes the outcomes of more than 12000 ward patients assessed for admission to critical care in 48 acute NHS hospitals. We observed a 90-day mortality of 30%. This is more than double that of similar unselected admissions even after we purposefully excluded patients with treatment limitations.[Scotland ref] More than a quarter (28%) of these deaths occured in the first 48 hours after the bedside assessment, and half (52%) of deaths in the seven days occured on the ward without admission to critical care.

Delay to admission was also common with just under a half (45%) of patients receiving critical care after the four hour target. This target is not ambitious by international standards.[Stelfox et al., 2012, #4003; Robert et al., 2012, #6549] The bedside decision strongly affected the odds of an early admission, but this decision was in turn affected by critical care capacity.

Although we only intended to use critical care occupancy as a tool to remove bias in estimating the effect of early admisson, the significant proportion of assessments that occur when the unit is full is an important finding itself. One in six patients studied (19%) were assessed when there fewer than two beds available, and one in twelve (8%) when the critical care unit was completely full. This risk is similar to recent reports from Canada and France, but markedly exceeds the 1% reported risk in the USA.[Robert et al., 2012, #6549; Stelfox et al., 2012, #4003; Chen et al., 2012, #12591]

Although decision making was affected by occupancy, the proportion of patients judged at the bedside to require critical care did not change. This implies that there exists a cohort of patients who are admitted when capacity is not an issue. Notably, since the proportion of deaths that occur on the ward, without critical care, increases when occupancy rises, then it is possible that these same patients die without a trial of critical care.

We would argue that the harm arising from such ‘missed’ trials of critical care is part of the reason we have observed benefit for early admission. It is not entirely possible in an observational study to separate the delivery of critical care from its timing. Any attempt to isolate a purely ‘delayed’ group of patients for the purposes of comparison automatically creates selection bias. In a randomised controlled trial (RCT), this risk is mitigated by stipulating an ‘intention to treat analysis’ (ITT). Therefore, although the future may unfold differently for the treated and the controls, there is no risk of this creating bias. A hypothetical RCT of early versus delayed admission would require those assigned to the delayed arm to be included in the analysis even if they died before admission. Absurdly, they would also require those who improved while waiting to be admitted. An observational study that excludes these two subgroups wil be biased in favour of the delayed arm if deaths are excluded, and against the controls if the improvers are excluded. These biases are present in majority of previous studies of early versus delayed admission to critical care.[Restrepo et al., 2010, #13; Renaud et al., 2009, #600; Chalfin et al., 2007, #612; Rapoport et al., 1990, #616]

The majority (59%) of patients in our study would be classed as improvers since they survived to 90-days without critical care. Clinical judgement, not random allocation, selects the most unwell patients into the early admission arm. Although such judgement is inevitably imperfect, we can see that the measured severity of illness is much greater in those admitted early. We can adjust for these measured variables, and remove the naïve association observed between early admission and harm. However, despite our best efforts, this risk adjustment is likely to be insufficient. We used the ICNARC physiology score because it captures a greater number of physiology dimensions than the NHS National Early Warning Score (NEWS), and is less dependent on treatment variables than the SOFA score. This is still not ideal because ICNARC physiology score was calibrated in a different population (those admitted to critical care), and even in that population, when combined with additional risk factors, its prediction is inevitably imperfect.[Harrison 2006] To try and capture some of the clinician’s end of the bed judgement, we have included the clinician’s recommendation for critical care as a proxy for some of these unmeasured factors. Since it is through this recommendation that a decision to admit is delivered, then adjustment for this removes explanatory power. We would have prefered to understand the complete effect of early admission, not just the effect that is independent of the clinician’s recommendation. More generally, it seems unlikely that summary data collected at the bedside by the researcher will ever completely include all factors that a clinician uses to allocate treatment.

Our estimate of decreased mortality with early admission instead comes from an instrumental variable (IV) analysis. This replaces the assumption of adequate risk adjustment with the assumption that occupancy only affects outcome through delaying or blocking critical care admission. We believe this is the more reasonable assumption, and others have relied upon it, or similar.[Shmueli 2005; Kim 2012; Pirracchio 2011] Its value is that it leaves only one route by which differences in outcome can be linked to occupancy, and that is through early rather than delayed or blocked admission to critical care. Unadjusted mortality is 31.6% for those assessed when there are no beds, 30.9% when there is just one bed, and 30.2% otherwise. The instrument can be thought of as natural randomisation event with imperfect treatment compliance. The proportions of patients admitted early increases as occupancy falls from 9%, to 15% to 22% respectively. The IV methodology makes allowance for this imperfect ‘compliance’, and scales up the observed mortality difference from 1–2% to around 10% albeit with a loss of precision (95%CI 0.4 to 19.8%.)

This finding does not stand in isolation. In our own study, we can see that occupancy has effects on intermediate steps in the causal pathway. Therefore, those patients assessed when the unit is full, are not only admitted late, but this delay is associated with a greater increase in physiological severity while waiting. In previous work, reducing exposure to critical care at the end of an admission by premature discharge rather than delayed admission has also been shown to increases mortality.[Rowan 1998] At a population level, expanding critical care capacity through an increase in funding during health service reforms is similarly assocaited with an improvement in outcomes.[Hutchings 2009] On a smaller scale (5 hospitals, 749 patients), the one previous study of early admission to critical care, that similarly avoided selection bias by completing follow-up for all potential admissions, also found benefit for early admission.[Simchen 2004]

Notably, this latter study defined early admission based on the location of care during the first 24 hours following the onset of critical illness. We used four hours as our definition on the recommendation of our steering group. This was based on current UK practice, mirrors recent guidelines[ICS], and given that we exclude deaths during the assessment, minimises the risk of death or improvement in the period from decision to delivery of critical care. We attempted to evaluate the specific effect of admission timing without the influence of location of care by repeating our analysis in the subgroup offered critical care. This reduced but did not eliminate from the controls the subgroup of patients never admitted from 80% to 28%, made little substantial difference to the effect estimate for early admission other than reducing its precision (95%CI -0.9%–34%).

The most obvious limitation to our study is that we must rely on some form of assumption in order to evaluate the effect of early admission. The threat to the assumption that occupancy only affects outcome through delaying or blocking admission principally comes from the concept that crowding inside the critical care unit may also be harmful. The evidence for this is conflicting[Gabler et al., 2013, #60215; Kahn et al., 2009, #12665], and even if true, would only apply to thoe admitted. We also need to acknowledge that in an observational study of this size, there are limitations in the quality of the data we record. We captured 94% of admissions to critical care in the study database. We have further tested our findings by raising and lowering our threshold for judging a month complete (see ESM). We can find no consistent difference in any of these results other than a fall in precision as the quality threshold increases, and the sample size inevitably falls.

Aspects of the study are stand independent to the problems above. Regardless, of the effect of early admission to critical care, we have identified a cohort of hospital patients at very high risk. This risk is heavily front loaded, but also sustained through the follow-up period. The bedside assessment is an effective but imperfect triage tool, as the mortality in those initially refused admission is high. The fact that a large number of these patients die on the ward without critical care makes us anxious since, we believe, the weight of evidence shows benefit for this expensive resource. Although decision making is influenced by occupancy, clinical recommendation is not which suggests that there is a cohort of patients who are only admitted when capacity is not an issue. Expanding critical care bed numbers would first and foremost benefit this group. There is also an opportunity to create a virtuous circle here. Earlier admission might lead to shorter admissions thereby improving flow through critical care and outcomes together. Identifying those patients who should be admitted early is already the number one priority for both clinicians and patients.[ref JICS, JLA] What we have contributed we hope, is firm evidence in support of this.

# References

# Electronic supplementary material

# References

# Electronic Supplementary Material

ESM text to go here

# Tables

## Table 1

# Figures

## Figure 1

Figure 1: Ward referrals assessed for eligibility at participating hospitals, reasons for exclusion, and admission timing following bedside assessment.