Disentangling the volume-outcome relationship in neonatal care in England*

Sam Watson †[2,1], Wiji Arulampalam[1], Stavros Petrou[2] and Neena Modi[3]

[1]Department of Economics, University of Warwick, Coventry
[2]Warwick Medical School, University of Warwick, Coventry
[3]Department of Medicine, Imperial College, London

November 30, 2012

PRELIMINARY WORK, PLEASE DO NOT CITE

Abstract: We estimate the effects of patient volume on the mortality of infants treated in neonatal units, taking into account different possible sources of bias. Patient or procedure volume has been demonstrated to be inversely associated with adverse clinical outcomes such as mortality. However, when estimating the effects of volume, two sources of bias may be present: i) the direction of causality between volume and outcome may be bi-directional, and ii) relevant unit level variables correlated with volume may be omitted. We use an instrumental variables strategy along with a wide range of different unit level variables and measures of volume to estimate the effects of volume in English neonatal units. We further add to the literature by using an underexplored and rich dataset, the National Neonatal Research Database (NNRD). We included 12,426 infants in our analysis from 127 different neonatal units. We find evidence of a reduction in the risk of mortality for the largest units but an increase in the risk of mortality for units providing a greater proportion of intensive care or with fewer reported cots relative to the caredays provided.

Keywords: Hospital volume, neonatal unit, learning by doing, omitted variables bias

JEL Classification Numbers: I12, I18, D20

*We are extremely grateful to the team at the Neonatal Data Analysis Unit for all of their assistance and advice and to thank Andrei Morgan for his work with the Unit Profile Survey. We would also like to thank the NESCOP group and Bliss for many valuable discussions about this research.

†s.i.watson@warwick.ac.uk

1 Introduction

A large number of empirical studies have documented an association between patient or procedure volume and clinical and economic outcomes (see Halm et al. (2002) for a review). Typically it is found that patient volume is inversely associated with adverse clinical outcomes such as mortality. This result has been found for a large number of different procedures (for example, Barker et al. (2011), Farley and Ozminkowski (1992), Gaynor et al. (2005), Luft et al. (1979), Tsai et al. (2006), Phibbs et al. (2007), Luft et al. (1987), Baker and Phibbs (2002)). These studies have had a large impact on healthcare policy, with many authors advocating a centralisation of health services (see Phibbs (2012) for a recent editorial). However, increasing volume *per se* may not necessarily improve outcomes; it is the attendant decreases in average costs or increased experience that might. Omitting relevant unit level variables will bias the estimated effects of volume on outcomes. Our aim is to estimate the effects of volume on outcomes of babies treated in neonatal units in England controlling for a wide range of associated and possibly correlated factors such as labour, resourcing, and complexity of care, as well as addressing the possible simultaneous relationships between volume and outcome.

The two principle hypotheses about how the volume-outcome relationship functions are 'practice makes perfect', which describes the causal effect of volume on outcomes, and 'selective referral', which describes the causal effect of outcomes on volume (Luft et al., 1987). The latter case arises when hospitals with superior outcomes attract a greater demand for their services.[1] The former hypothesis may be decomposed into the effects of economies of scale and the effects of learning by doing. Economies of scale could influence quality by leading to an increase in labour and capital inputs. And, the effects of learning by doing may be affected by the marginal returns to experience; the learning curve may be 'steeper' for more complex cases. We may therefore expect there to be different thresholds of volume required to achieve high quality care for different patient groups.

In this study we examine patient volume in neonatal units and outcomes of babies treated within those units. Our objective is to identify the causal effect of volume on outcomes. However, given the possible simulataneous relationship between volume and outcome we use an instrumental variables strategy using the characteristics of the mother's nearest hospital along with the distance from that hospital as instruments for the characteristics of the hospital providing treatment. These instruments exploit the strong, exogenous preference for mothers to use their nearest hospital.

We study neonatal care for several reasons. In recent years, a growing body of research has shown the importance of early life influences on human capital development. The health of an in-

[1] In the UK this effect may be somewhat limited by the organisation of the National Health Service, however, as we will describe, we study a system which has been designed to transfer the sickest patients to the largest hospitals. Hence selective referral may be a significant concern in our study.

dividual in the first few days of life has been shown to have important effects on health outcomes, labour market participation, educational performance, and earnings (Almond and Mazumder, 2011, Black et al., 2007). Providing high quality health care services for infants is therefore vital and may have large welfare implications over the course of a lifetime both through the large number of life years saved and also by reducing socioeconomic disparities. Infants admitted onto neonatal units represent a sizeable population; approximately 10% of all infants receive some care on a neonatal unit.[2] It is important that policy decisions regarding these units and the care they provide are supported by empirical evidence and, particularly in the case of the National Health Service (NHS) in England, demonstrably cost-effective.

A number of empirical studies have demonstrated an association between patient volume and clinical outcomes in neonatal care (e.g. Baker and Phibbs (2002), Chung et al. (2010, 2011), Cifuentes et al. (2002), Phibbs et al. (2007)). However, these studies have not necessarily demonstrated the causal effect of volume on outcomes. Indeed, only a handful of studies of the hundreds published have arguably demonstrated a causal relationship (Barker et al. (2011), Farley and Ozminkowski (1992), Gaynor et al. (2005), Luft et al. (1979), Tsai et al. (2006), see Halm et al. (2002) for a review). Furthermore, none of the aforementioned studies have taken into account the possible bias caused by omitted variables.

We use a rich and under-explored dataset, the National Neonatal Research Database (NNRD), currently held by the Neonatal Data Analysis Unit (NDAU) at Imperial College, London. The NNRD contains the electronic patient records of all infants admitted to 162 of 173 neonatal units in England dating back to 2006. We also use data from the Unit Profile Survey (UPS), a survey conducted of neonatal units in England in 2011 on resourcing and patient volume[3].

The rest of this paper is organised as follows. Section 2 summarises the background research underpinning the study both in general and in neonatal care. Section 3 details the data sources and sample selection procedures. Section 4 explains the methods; section 5 presents the results, and section 6 presents some conclusions.

2 Background

2.1 Patient volume and outcomes

Many hundreds of studies have been published on the volume outcome-relationship over the last 25 years. Predominantly, these studies have identified a positive association between patient

[2]In 2011 there were 688,120 live births and over 68,000 discharges from neonatal units (Office of National Statistics, 2012, Royal College of Pediatrics and Child Health, 2012)

[3]The UPS is a follow up survey to the previously conducted EPICure survey on neonatal unit resourcing and staffing (Costeloe et al., 2000)

volume and clinical outcomes (Halm et al., 2002). However, only a handful of these studies have acknowledged the potential simultaneous nature of volume and outcome. Cross-sectional studies that do, have used various econometric strategies for accounting for this simultaneous relationship. In particular, authors have used an instrumental variables technique to address the endogeneity issue caused by selective referrals. Selective referrals may arise due to patients choosing hospitals with lower mortality rates or higher reputations or doctors transferring sicker patients to larger hospitals. Authors have often used market size or patient distance to hospital as instruments for hospital volume. Many of these studies have found that the estimated causal effects of volume on outcomes diminishes and occasionally becomes statistically insignificant when instruments are used (Barker et al., 2011, Farley and Ozminkowski, 1992, Gaynor et al., 2005, Luft et al., 1979, Tsai et al., 2006).

In a recent study of cardiac revascularisation in US hospitals, Barker et al. (2011) sought to determine whether speciality cardiac hospitals had reduced mortality compared to general hospitals. The authors used patient distance to hospitals to determine expected volume and found evidence of significant simultaneity bias. Gaynor et al. (2005) used a similar method to examine the effects of volume on outcomes for coronary artery bypass graft (CABG) surgery in US hospitals. The authors included contemporaneous and lagged measures of volume to take account of learning over time but also to capture an effect of forgetting. They found that the parameter estimates for lagged volume were insignificant indicating a strong effect of forgetting and suggesting that the volume-outcome relationship, for CABG at least, is solely contemporaneous.

The aformentioned studies examine one procedure to identify the mean treatment effect of volume. However, it is possible that some types of patients may experience a larger benefit from volume while others experience no benefit at all. The treatment effect may be a function of patient or procedure characteristics.

We are aware of only two studies that both attempt to eliminate the simultaneity bias and to look at different procedures and patient groups in the same hospital (Farley and Ozminkowski, 1992, Luft et al., 1979). Both of these studies found that 'practice makes perfect' was only present in some of the procedures examined. Farley and Ozminkowski (1992) used panel data on up to 432 hospitals (depending on the procedure) over eight years to explore the volume-outcome relationship for five procedures. They observed a significant relationship between volume and outcome for all five procedures, although this relationship only remained for three after using instrumental variables to control for selective referral. Similarly, Luft et al. (1979) examined 12 procedures of varying complexity in 1,498 hospitals and found that only some exhibited a 'practice makes perfect' relationship. We are not aware of any studies that extend volume-outcome analyses by including controls for labour and capital inputs as well.

2.2 Research in neonatal care

Neonatal care has seen an increasing body of research devoted to understanding the link between unit characteristics and patient outcomes. This patient group is a vulnerable one, neonatal mortality comprised 69% of the overall infant mortality rate in England and Wales in 2010 (Office of National Statistics, 2011). Moreover, spending on neonatal healthcare, particularly on capital inputs, has expanded in recent years. Many studies of neonates focus on the most vulnerable groups, in particular babies who are classified as very low birth weight (VLBW, birth weight less than 1500g). This group of babies is generally the most severely ill and requires the highest cost care, but only represents a small proportion of the overall patient population.

In neonatal care, patient volume, delineated either in terms of the number of babies treated in a hospital over a fixed length of time or the number of cots, has been of high interest. In the United States, a number of papers have remarked on the trend towards deregionalisation of neonatal services and its implications (Baker and Phibbs, 2002, Phibbs et al., 2007). In the United Kingdom, studies demonstrating a positive effect of patient volume have been cited in documents leading to the formation of cooperative neonatal networks (Department of Health, 2003). However, none of these studies have demonstrated a causal effect of volume on mortality or indeed any other outcome.

A widely cited paper on the topic from the United States by Phibbs et al. (2007) found that VLBW babies born in the highest intensity units, each with an annual admission rate of 51-100, were 19% more likely to die than those born in the highest intensity units with an annual VLBW admission rate of greater than 100. Similar results have been presented by other studies from the US (Rogowski et al., 2004, Chung et al., 2011, 2010, Cifuentes et al., 2002). However, these studies present only correlations and may be biased for the reasons cited in the introduction.

The studies described in this section have examined the average treatment effect of volume on outcomes. However, as we explain, we expect there to be a heterogeneous effect of volume depending on patient severity. A recent study using Californian brith records attempted to identify a heterogenous treatment effect of length of stay on infants treated in neonatal units and did find a much stronger effect for more severely ill babies (Evans and Garthwaite, 2012).

The literature from the United Kingdom on patient volume has been a lot sparser. Only two studies that we are aware of have estimated the effects of patient volume on outcomes in English neonatal units, neither of which found a statistically significant effect (Tucker, 2002, Hamilton et al., 2007)[4]. However, Tucker (2002) did find that babies admitted to units at higher occupancy were more likely to die, whilst Hamilton et al. (2007) demonstrated an inverse association between staffing and neonatal mortality. As we have discussed, if occupancy or staffing were correlated

[4]Both of these studies used data from the survey preceding the one used in this study.

with volume and omitted from an analysis of volume then the estimated effect of volume would be biased.

2.3 Neonatal care in England

In 2003, the Department of Health published the *National Strategy for Improvement*, the result of a national review into the organisation of neonatal care services. This document found that neonatal care was widely dispersed and that the pattern of transfers could be improved for babies, parents and staff. Furthermore, the report found that the national capacity for neonatal care was inadequate. It was recommended that managed clinical networks (MCNs) be formed, rather than major centralisation, as equity was an important concern; intensive care stays are often long and parents would be burdened with long journeys away from home. Additionally, only the sickest babies which represent the minority would necessarily need to be transferred.

Within these networks, units are designated different levels depending on the care provided.[5] The highest intensity of the three designated levels is a neonatal intensive care unit (NICU; level 3), the next highest is a high dependency unit (HDU; level 2), and the lowest intensity unit is a special care baby unit (SCBU; level 1). The levels generally correspond to the ability of that unit to provide the three categories of care for babies admitted onto neonatal units; these categories of care are generally, although not exclusively, defined on the basis of the type of respiratory support a baby receives. A baby's requirement for care is strongly correlated with his or her gestational age and birth weight along with other clinical factors (Tucker and McGuire, 2004). We provide further supporting evidence for this in section 5. The unit level designation is therefore strongly correlated with the facilities it provides and thus its capital inputs - we use this designation to control in part for capital inputs.

3 Data

3.1 Data sources

For our analysis we use two sources of data. The first is the National Neonatal Research Database (NNRD). The NNRD was created by the Neonatal Data Analysis Unit (NDAU), a research unit based at Imperial College, London, and was established in 2006, using individual patient level records of infants treated within neonatal units in England. The NDAU holds national research ethics committee approval to create this database. The data are psuedo-anonymised by removing patient and maternal identifiers and encrypting the NHS number of each patient.

[5]Each hospital only has one unit.

The data include a vast range of variables including static descriptive variables captured once per baby, such as birth weight and gestational age at birth, episodically, such as episodes of infection and other clinical outcomes, and daily items such as treatments and procedures as well as level of care. However, data quality and number of entries that are fully completed vary by field.

The NNRD is a rich source of information. Data are available from the inception of the NNRD in 2006, however the quality of some aspects of this data is questionable at this time. Complete data were available for analysis from between January 2006-August 2012. The number of units contributing to NDAU increased year on year and ranged from 123 in 2007 to 153 in 2011 of 165 units in total in England.

The second source of data used here was the Unit Profile Survey 2011 (UPS). This survey was posted in November 2011 to all English neonatal units to obtain cross sectional data on staffing, resourcing, and organisation; it was conducted as part of the Neonatal Economic, Staffing and Clinical Outcomes Project (NESCOP). The response rate to the survey was 95% (156 units)[6].

3.2 Analysis sample

As the staffing cross-sectional survey was taken in November 2011 we restrict our sample to babies born and discharged between January 1st 2011 and the most recently available date, which was 31st August 2012.

From the NNRD we extract babies born at or less than 32 weeks gestation. Babies born above this threshold are admitted for a variety of reasons and are likely to have a different set of variables determining admission and outcomes than more preterm infants (McIntire and Leveno, 2008, Bastek et al., 2008). Also, our outcome in this study, mortality, is strongly related to gestational age at birth and is rare in babies born above 32 weeks gestation. We exclude babies for whom we do not observe all the covariates included in the models.

The UPS survey was conducted in November 2011. For labour inputs the UPS provided details on the number of whole time equivalent (WTE) nurses per unit and the number of consultants with at least 50% of their time dedicated to neonatal care. However, since the composition of patients differs by hospital the amount of work required differs by unit. We would expect a unit with the same number of staff to perform better if the average patient required less complex and time consuming care. Therefore, it is important to adust our measure of labour for the work required. The British Association of Perinatal Medicine (BAPM) defines an 'adequate' level of nursing to be a function of $4*IC+2*HD+SC$[7], where IC is the number of intensive care babies, HD the number

[6]This study is a follow up study a previously conducted survey on neonatal unit resourcing and staffing (Costeloe et al., 2000). Studies to use data from this previous survey include Tucker (2002), Hamilton et al. (2007)

[7]Specifically, the recommendation is a 1:1 nurse to baby ratio for babies requiring intensive care, 1:2 ratio for high dependency, and 1:4 ratio for special care

of high dependency care babies, and SC the number of special care babies (British Association of Perinatal Medicine, 2001). We maintain these ratios of staff required to care required and define the workload in terms of the annual number of care days at each level to reflect actual work each year and we multiply by 365 to reflect a daily ratio.

Our outcome variable is neonatal mortality defined as mortality within 28 days post birth. This is a common outcome in studies of neonatal care (e.g. Phibbs et al. (2007), Cifuentes et al. (2002), Chung et al. (2010)); moreover, the vast majority of all deaths on neonatal units occur in the first 28 days.

4 Method

We estimate the following model, where the unit of observation is baby, i, treated in year, t:

$$y_{it} = \alpha + X_{it}'\beta + Z_{it}'\gamma + \alpha_t + u_{it} \tag{1}$$

Where y_{it} is scalar variable equal to 1 if the infant died within 28 days and 0 otherwise, X_{it} is a vector of exogenous baby covariates, Z_{it} is a vector of potentially endogenous unit covariates, α_t are time dummies for each year, and β and γ are vectors of parameters to be estimated.

We estimate (1) by OLS, treating the Z_{it} as exogenous. However, there may be some unobserved heterogeneity between the babies - this unobserved difference will be absorbed into the error term such that $Cov(z_{it}, u_{it}) \neq 0$. The sign of the covariance depends on the unit level variable, for example, we expect that $Cov(z_{1it}, u_{it}) > 0$, where z_{1it} is the unit volume, as (unobservable) sicker babies are sent to higher volume hospitals. This would mean estimates of the effects of volume would be biased upwards when estimated by OLS. We address this endogeneity by estimating equation (1) by instrumental variables estimation (IV) using a set of instruments. In Z_{it} we include the interaction of unit volume and gestational age, this interaction term is also potentially endogenous thus we need instruments for it. We create additional instruments by interacting gestational age with all our instrumental variables as well as including interactions of volume with the other instruments; this gives us 24 instruments. This method is valid provided these new instruments are (partially) correlated with the endogenous variables, Z_{it}, and not with u_{it}.

4.1 Covariates

The baby variables included are: gestational age at birth, gestational age squared[8], small for gestational age status[9], an indicator for multiple birth, indicators for whether a full or partial course of antenatal steroids was administered, and a dummy variable for male gender. These covariates are those commonly used in clinical studies of neonatal health and are included in the mortality models routinely used by clinicians as they have been shown to predict survival well (Parry et al., 2003, Draper et al., 2003, Tyson et al., 2008).

4.1.1 Unit level variables

The upper panel of table 1 shows all the unit level variables included in the analysis. We considered a number of different measures of volume which are shown in the lower panel of table 1. We use VPT caredays in the main analysis but re-estimate the models using each measure in a subsequent analysis. We define workload as shown in table 1 since these are the ratios used by the British Association of Perinatal Medicine (BAPM) when they define an 'adequate' level of nursing. Our intention is to incoporate proxies for each of the independent variables in **??** and so we consider the level 3 dummies, the cot ratio and the proportion of intensive care to be proxies for capital inputs within a unit. Level 3 units are specifically designated as units that provide the highest intensity care and so should be appropriately equipped. Similarly, units that are more well equipped may have a greater number of cots relative to the amount of care provided. Note that our cot ratio variable is the reciprocal of unit cot occupancy. Our variables of nurses and consultants provide measures of labour inputs. We consider the remaining effect of volume to channel through experience; for this reason the interaction of experience and gestational age is included to capture a 'learning curve' effect.

5 Results

Overall, 12,426 infants, treated in 127 different neonatal units were included in the study. Of these we had data on the nearest unit for 10,984 babies. Descriptive statistics of the sample are shown in table 2

[8]The relationship between gestational age and neonatal outcomes is not linear and so, as with some other studies, a quadratic term is included.

[9]Small for gestational age is commonly defined as being in the bottom 10% of birth weights for a gestational age week conditional on gender and whether the infant was from a singleton or multiple birth. The values were taken from birth weight centile charts in Bonellie et al. (2008)

Table 1: Unit level variable definitions

Variable	Definition
ln(volume)	The logarithm of the volume measure.
ln(volume) ∗ gest.age	The interaction of the logarithm of volume and gestational age
Level 3, without	Dummy variable equal to 1 for a level 3 units without specialist surgery or cardiac services
Level 3, with surgery	Dummy variable equal to 1 for a level 3 units with specialist surgery but not cardiac services
Level 3, with surgery and cardiac	Dummy variable equal to 1 for a level 3 units with specialist surgery and cardiac services
Cot ratio	Total number of cots multiplied by 365 divided by the total number of care days
PropIC	Annual number of intensive care days divided by total number of care days multiplied by 100
nurse/workload	Total number of WTE nurses multiplied by 365 divided by the workload
cons/workload	Total number of WTE consultants[a] multiplied by 365 divided by the workload
VPT caredays	Total number of caredays in year provided to babies born at or below 32 weeks gestation
Total caredays	Total number of caredays that year provided to any baby
VPT babies	Total number of unique babies born at or less than 32 weeks gestation admitted within 1 day of birth
IC caredays	Total number of intensive care days provided that year
Workload	Equal to $4 * IC + 2 * HD + SC$ where IC, HD, and SC are annual intensive care, high dependency care, and special care days.

[a] Consultants with at least 50% of their time dedicated to neonatal care
WTE=whole time equivalent

Table 2: Observed characteristics by level of nearest hospital

Variable	
N	12426
Gestational age at birth	29.38 (2.49)
SGA	10.37%
Antenatal steroids	68.52%
Multiple birth	21.32%
Male	50.06
Neonatal mortality	5.00%
Treated in level 3 unit	52.58%
N	127
Level 3	29.92%
Annual VPT caredays	2646.27 (1790.90)
Annual VPT babies[a]	307.6
Reported number of cots	20.53 (10.45)
% intensive care[b]	12.25 (8.78)
WTE nurses	43.03 (32.96)
Consultants[c]	3.36 (3.18)

[a] Babies admitted within 24 hours of birth
[b] Annual number of intensive care days as a proportion of all care days
[c] Consultants dedicating at least 50% of their time to neonatal care
WTE=whole time equivalent, VPT=very preterm, babies born at or less than 32 weeks gestation
Upper panel shows baby level statistics, the lower panel shows unit level statistics
Statistics are mean (sd) unless otherwise stated.

5.1 Instrument validity

We first validate the instruments we use to control for potential unobserved heterogeneity and simultaneity in the babies. Our instruments are the set of unit level variables for the unit nearest to the mother's place of residence. We don't observe the exact place of residence for each mother, rather we have data on their Lower Super Output Area (LSOA)[10]; distances are then calculated from the population weighted centroid of each LSOA[11]. We also include the distance as an instrument. Since one of our potentially endogenous unit level variables is an interaction between an endogenous and an exogenous variable, gestational age at birth, we include as instruments the set of interactions of nearest unit level variables interacted with gestational age at birth. The validity of our instrument relies on the assumption that mothers do not choose their place of residence on the basis of the nearest neonatal unit, which we do not believe is a strong assumption.

As we have more instruments than endogenous variables we can test whether the null hypothesis that our second stage errors are uncorrelated with our instruments with a Hansen Sargan test. The p-value from this test is reported for each outcome in table 6 along with the estimated coefficients. In each case we reject the null hypothesis that the model is under-identified and in each case the p-value from an F-test on the first stage regression coefficients strongly rejects ($p<0.0001$) the null that the instruments are irrelevant.

We further validate our instruments by examining whether babies are randomly distributed by instrument categories. We should not expect to see a statistically significant difference in the observed baby level covariates by instrument, but we should expect significant variation in the unit level variables. Table 3 shows observed characteristics of each baby by whether the nearest unit is level three or not, we also present p-values of tests examining whether there is a significant difference between the two groups.

5.2 Gestational age and complexity

We use a baby's gestational age as a proxy for complexity, we justify this as gestational age at birth is strongly correlated with the probability of mortality and the level of care required. In our sample, 23.8% of babies born at 24 weeks gestation died within 28 days of birth whereas this figure was only 3.3% for babies born at 28 weeks gestation. Figures 1(a) and 1(b) show respectively the proportion of babies receiving at least three days of intensive care and the proportion who received an arterial line by gestational age.[12] The dip at 23-25 weeks in intensive care in figure 1A reflects

[10]LSOAs are geographical areas designed to facilitate the reporting of geographical statistics in the UK. In England there are 32,844 LSOAs, all with a population between 1,000 and 3,000.

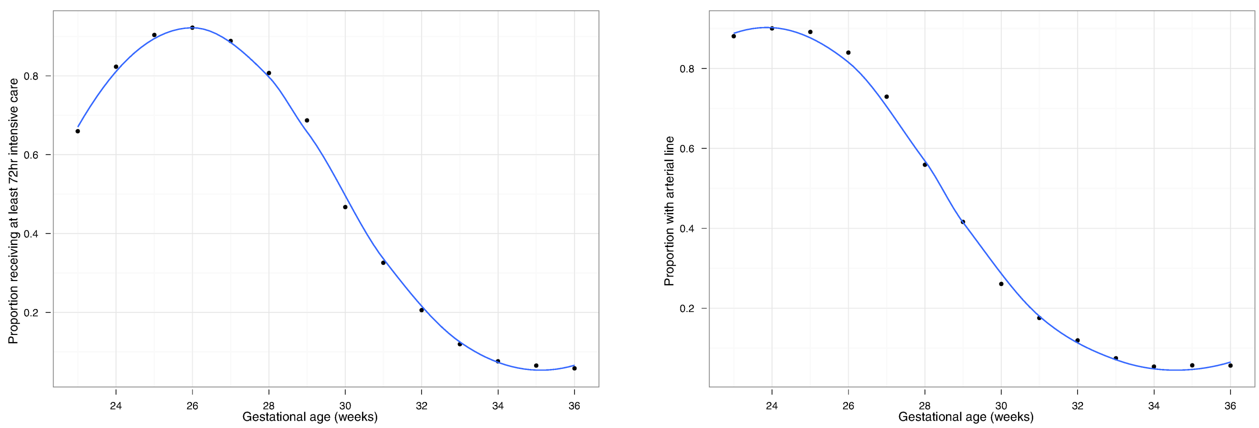
[11]The Office of National Statistics provide the eastings and northings of LSOA population weighted centroids from which we used the nearest neighbour algorithm to define straight line distances to the nearest hospitals

[12]An arterial line is often used in intensive or high dependency care settings when frequent blood pressure monitoring and blood sampling is required and also for patients who need an exchange transfusion. They are normally only

Table 3: Observed characteristics by level of nearest hospital

	Non-Level 3	Level 3	p-value[a]
N	4346	8080	
Gestational age (mean (sd))	29.35(2.50)	29.44(2.48)	0.07
S.G.A.	10.30%	10.49%	0.73
Gender (males)	50.00%	50.18%	0.85
Antenatal steroids	69.44%	66.80%	0.06
Multiple birth	22.07%	19.93%	0.01
Neonatal mortality	4.93%	5.13%	0.62
Treated in lev 3 at 24 hrs	34.22%	86.70%	0.00

[a] We used a t-test for continuous variables and chi sq test for categorical variables



(a) Proportion of babies by gestational age week at birth receiving at least 72 hours of intensive care

(b) Proportion of babies by gestational age week at birth with arterial line

Figure 1: Gestational age and complexity of care

the high level of mortality prior to 3 days for this group.

5.3 Main results

5.3.1 OLS

Table 4 shows the estimates of γ by OLS. We present OLS results first as we can compare these to later 2SLS results, but also to draw comparisons with previous studies. We find that larger volume hospitals are associated with a reduced risk of mortality, which is in line with other studies of neonatal care (see section 2.2 for a brief review). The size of the estimate of the coefficient for log volume is increasing across columns (1)-(4) in table 4, which suggests that volume is correlated with other unit variables and is biased upwards by their omission. For example, if the effect of

used for short periods of acute or critical illness Perrin (2008)

Table 4: OLS estimated coefficients for unit level variables

	(1)	(2)	(3)	(4)
ln(*volume*)	-0.00964*	-0.0212***	-0.0256***	-0.0268***
	(0.00389)	(0.00584)	(0.00640)	(0.00616)
Level 3, without surgical or cardiac		-0.00402	-0.00205	0.00196
		(0.00946)	(0.00937)	(0.00883)
Level 3, with surgical		-0.00438	0.00119	0.00676
		(0.0118)	(0.0119)	(0.0117)
Level 3, with surgical and cardiac		0.00792	0.0130	0.0212
		(0.0170)	(0.0166)	(0.0162)
Cot ratio		-0.0113	-0.0104	0.000981
		(0.00894)	(0.00889)	(0.00799)
PropIC		0.000925	0.000708	0.000241
		(0.000661)	(0.000663)	(0.000636)
$ln(volume) * gest.age$			0.00742**	0.00735**
			(0.00240)	(0.00239)
$nurses/workload$				-0.0150**
				(0.00517)
$consultants/$ $workload$				0.00773
				(0.0191)
Observations	12426	12426	12426	12426
Hospital units	127	127	127	127
R squared	0.1188	0.1196	0.1216	0.1227
Root MSE	0.2046	0.2046	0.2043	0.2042
$H_0 : \gamma = 0$	0.0145	0.0200	0.0158	0.0050

Cluster adjusted standard errors in parenthesis
† $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The models were adjusted for gestational age, gestational age sq., small for gestational age, multiple birth, gender, antenatal steroids and year dummies. Gestational age is centered around its mean of 29.55.

increased resources was to reduce the risk of mortality and resourcing was inversely correlated with mortality then we may see such bias. Table 5 shows the correlation between different dependent variables for units included in our sample. They show that higher volume units have smaller values for cot ratios, nurse ratios and consultant ratios and a greater proportion of their care devoted to intensive care. This would explain the bias seen in columns (1) to (4) of table 4.

The estimated effect of volume is two and a half times greater once other unit level variables have been added showing evidence of significant omitted variable bias. However, these estimates themselves may be biased due to endogeneity of the unit level variables.

Table 5: Correlation between different independent, unit-level variables

	Volume	Cot ratio	Nurse ratio	Cons. ratio	Prop. IC
Volume	1.00				
Cot ratio	-0.40	1.00			
Nurse ratio	-0.53	0.64	1.00		
Cons. ratio	-0.05	0.15	0.20	1.00	
Prop. IC	0.72	-0.09	-0.29	0.00	1.00

5.3.2 Instrumental Variables

Table 6 summarises the results from the models estimated by IV. The IV estimates for volume are larger in magnitude than the OLS estimates, which indicates that the OLS results were biased upwards as expected. This suggests that infants who were unobservably more likely to die were transferred to higher volume hospitals. We also observed a change in other coefficients. In particular, the coefficient for the ratio of cots to mean daily patients is significant which suggests that a decrease in the average occupancy can reduce mortality. We also find that a one percentage point increase in the proportion of intensive care days to total care days leads to an increase in the risk of mortality by 0.166 percentage points *ceteris parabus*. In our sample, the proportion of intensive care days ranges from under 1% to over 30%. However, the effect of intensive care is balanced by the effect of increased volume.

With regards to staffing we find the IV estimates differ from the OLS estimates. Specifically, the estimate of the effect of nurses is approximately half the size and becomes insignificant. This suggests that unobservable sicker babies are more likely to be sent to hospitals with fewer nurses relative to the workload. As table 5 shows, this ratio is negatively correlated with volume. The opposite effect is seen with the ratio of consultants to workload, the IV estimated coefficient is now negative but not statistically significant.

As with OLS, we find evidence of significant omitted variables bias. Once all the other unit level variables are taken into account the magnitude of the effect of volume is estimated to be over three times greater.

5.3.3 Marginal effects

The estimated marginal effects of volume and bootstrapped standard errors are summarised in table 7. This shows a clear effect of volume reducing the risk of mortality. For 24 week gestation babies, a doubling in volume is expected to lead to a 6.29 percentage point decrease in the risk of mortality, or 24% of the sample mean predicted probability of mortality for this group. As a proportion of the sample mean for different groups the marginal effect of volume is increasing, for

Table 6: IV estimated coefficients for unit level variables for neonatal mortality

	(1)	(2)	(3)	(4)
ln(volume)	-0.00992*	-0.0286***	-0.0335***	-0.0344***
	(0.00393)	(0.00667)	(0.00866)	(0.00861)
Level 3, without surgical or cardiac		-0.00304	-0.00194	0.000739
		(0.0112)	(0.0111)	(0.0112)
Level 3, with surgical		-0.0266	-0.0231	-0.0193
		(0.0140)	(0.0140)	(0.0144)
Level 3, with surgical and cardiac		-0.00394	0.000853	0.00383
		(0.0184)	(0.0186)	(0.0195)
Cot ratio		-0.0364*	-0.0369*	-0.0328*
		(0.0146)	(0.0144)	(0.0152)
PropIC		0.00210**	0.00185*	0.00166*
		(0.000790)	(0.000796)	(0.000813)
$ln(volume) * gest.age$			0.00479	0.00513
			(0.00462)	(0.00464)
$nurses/workload$				-0.00571
				(0.00501)
$consultants/workload$				-0.0287
				(0.0314)
Observations	10984	10984	10984	10984
Hansen test p-val[a]	0.1114	0.1890	0.3726	0.3509
$H_0 : \gamma = 0$[b]	0.0720	0.0003	0.0006	0.0005
Uncentered R-squared	0.1654	0.1636	0.1654	0.1659
Root MSE	0.2034	0.2035	0.2033	0.2033
Number of instruments	19	21	21	21

Cluster adjusted standard errors in parenthesis

† $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The models were adjusted for gestational age, gestational age sq., small for gestational age, multiple birth, gender, antenatal steroids and year dummies. Gestational age is centered around its mean of 29.55.

[a] The Hansen test has the null hypothesis that the residuals from the structural equation are uncorrelated with the instruments

[b] This is the p-value for an F-test that the instrumented variables are jointly different from zero.

Table 7: Estimated marginal effect of log volume

Gest. age (weeks)	Marginal effect (ppc)	Gest. age (weeks)	Marginal effect (ppc)
24	-6.29	30	-3.21***
	(3.87)		(0.95)
26	-5.26†	32	-2.19*
	(2.77)		(1.04)
28	-4.24*		
	(1.82)		

Bootstrapped standard errors in parenthesis
Marginal effects and standard errors are multiplied by 100.
† $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

example a 2.19 percentage point reduction in the risk of mortality for 32 week gestation babies is 83% of the sample mean. This suggests that the causes of death among higher gestational age babies are more preventable by quality care than the causes of death among the earliest gestation babies.

5.4 Volume measures

Table 8 shows 2SLS estimates using different measures of volume, we repeat the results using our original measure of volume in column (1). The definitions of the different measures of volume are shown in table 1. With the exception of the dummies for unit status, all the coefficients have the same sign regardless of the measure of volume. The greatest estimated effect is seen with our measure of VPT care days. The effect of cot ratio is statistically significant in columns (1), (4), and (5) and of a similar magnitude in all columns. Also, parameter estimates for staffing are all negative although statistically insignificant. The large standard errors for staffing estimates may be due either to a poorly fitting first stage or low variance in the observed values of staffing.

Examining different measures of volume provides further insight into the volume-outcome relationship. While all the measures of volume are highly correlated, as shown in table 9, the estimated marginal effect does differ somewhat. For example, a doubling in the number of VPT care days is expected to lead to a change in the risk of mortality for a 28 week gestation infant of -4.24 percentage points, whereas a doubling in total care days has a predicted effect of -4.89 percentage points. The change in the marginal effect of volume with gestational age is greater for total care days than for VPT care days.

The estimated volume coefficient for number of very preterm admissions is negative but not significant as shown in column (3) of table 8. This suggests that the effect of volume is due to the amount of care provided rather than absolute number of babies. This lends support to the hypothesis that the estimated volume coefficient, after controlling for labour and capital inputs, is

Table 8: 2SLS estimated coefficients for unit level variables for neonatal mortality

	(1) VPT caredays	(2) Total caredays	(3) Number of VPT babies	(4) IC caredays	(5) Workload
ln(volume)	-0.0344***	-0.0320**	-0.00766	-0.0166	-0.0284**
	(0.00861)	(0.0109)	(0.0172)	(0.0101)	(0.0110)
Level 3, without	0.000739	-0.00222	-0.00540	-0.00575	-0.00380
	(0.0112)	(0.0109)	(0.0111)	(0.0111)	(0.0110)
Level 3, with surgical	-0.0193	-0.0161	-0.0264	-0.0336†	-0.0213
	(0.0144)	(0.0154)	(0.0169)	(0.0175)	(0.0164)
Level 3, with surgical and cardiac	0.00383	0.00900	-0.00285	-0.0114	0.00307
	(0.0195)	(0.0197)	(0.0217)	(0.0214)	(0.0207)
Cot ratio	-0.0328*	-0.0349	-0.0193	-0.0476*	-0.0409*
	(0.0152)	(0.0182)	(0.0152)	(0.0210)	(0.0199)
$ln(volume) * gest.age$	0.00513	0.0109*	0.0104	0.00170	0.00681
	(0.00464)	(0.00502)	(0.00745)	(0.00353)	(0.00456)
$nurses/workload$	-0.00571	-0.00704	-0.00182	-0.00696	-0.00765
	(0.00501)	(0.00507)	(0.00582)	(0.00546)	(0.00517)
$consultants/workload$	-0.0287	-0.0503	-0.0438	-0.0217	-0.0475
	(0.0314)	(0.0381)	(0.0380)	(0.0334)	(0.0381)
PropIC	0.00166*	0.000815	0.000452	0.00223†	0.00151
	(0.000813)	(0.000800)	(0.000810)	(0.00119)	(0.000924)
Observations	10984	10984	10984	10984	10984
Hansen test p-val[a]	0.4158	0.3215	0.2067	0.4693	0.4435
Uncentered R-squared	0.1659	0.1669	0.1652	0.1630	0.1659
Root MSE	0.2033	0.2031	0.2033	0.2036	0.2032

Cluster adjusted standard errors in parenthesis
† $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The models were adjusted for gestational age, gestational age sq., small for gestational age, multiple birth, gender, antenatal steroids and year dummies. Gestational age is centered around its mean of 29.55.
[a] The Hansen test has the null hypothesis that the residuals from the structural equation are uncorrelated with the instruments
All volume measures are the log of the measure reported and are annual.
Volume measures are as defined in table 1.

picking up an effect of experience.

6 Discussion and future direction of research

In this section we highlight some of the weaknesses of the study and detail the future direction of the research including possible re-estimation strategies.

We have specific concerns regarding the validity of some of the data collected by the Unit Profile Survey and in how the units completing the survey interpreted the questions. In particular, the number of cots reported by each unit is not necessarily representative of the number of cots in active use in the unit. Whether a cot is ‘closed’ or not is dependant on a number of other factors such as staffing. However, we expect that the reported number of cots is well correlated with the number of cots in use.

We intend to improve and expand on this analysis in a number of different ways:

Table 9: Correlation between different measures of volume

	VPT caredays	Total caredays	VPT admissions	IC caredays	Workload
VPT caredays	1.0000				
Total caredays	0.8886	1.0000			
VPT admissions	0.6942	0.8479	1.0000		
IC caredays	0.8081	0.9110	0.7773	1.0000	
Workload	0.8811	0.9766	0.8171	0.9744	1.0000

- Repeat the analysis using only NDAU data from 2009-2012 to eliminate potential problems with UPS data.
- Estimate our model with unit fixed effects as we can use multiple years of unit volume and activity data. This will also allow us to observe if our unit level variables are adequate. Furthermore, this may eliminate time invariant unit unobserved characteristics confounding our analysis. However, there is not much variation over time in volume.
- Re-estimate the model with different gestational age groups to allow all unit level estimated coefficients to vary by complexity.
- Estimate the model as a least squares dummy variable model.

We welcome any discussion on these topics.

References

Almond, D. and Mazumder, B. (2011). Health Capital and the Prenatal Environment: The Effect of Ramadan Observance During Pregnancy. *American Economic Journal: Applied Economics*, 3(4):56–85.

Baker, L. C. and Phibbs, C. S. (2002). Managed care, technology adoption, and health care: the adoption of neonatal intensive care. *The Rand journal of economics*, 33(3):524–48.

Barker, D., Rosenthal, G., and Cram, P. (2011). Simultaneous relationships between procedure volume and mortality: do they bias studies of mortality at specialty hospitals? *Health economics*, 20(5):505–18.

Bastek, J. A., Sammel, M. D., Paré, E., Srinivas, S. K., Posencheg, M. A., and Elovitz, M. A. (2008). Adverse neonatal outcomes: examining the risks between preterm, late preterm, and term infants. *American journal of obstetrics and gynecology*, 199(4):367.e1–8.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2007). From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes. *The Quarterly Journal of Economics*, 122(1):409–439.

Bonellie, S., Chalmers, J., Gray, R., Greer, I., Jarvis, S., and Williams, C. (2008). Centile charts for birthweight for gestational age for Scottish singleton births. *BMC pregnancy and childbirth*, 8:5.

British Association of Perinatal Medicine (2001). Standards for hospitals providing neonatal intensive and high dependency care. Technical report, BAPM, London.

Chung, J. H., Phibbs, C. S., Boscardin, W. J., Kominski, G. F., Ortega, a. N., Gregory, K. D., and Needleman, J. (2011). Examining the effect of hospital-level factors on mortality of very low birth weight infants using multilevel modeling. *Journal of perinatology : official journal of the California Perinatal Association*, 31(12):770–5.

Chung, J. H., Phibbs, C. S., Boscardin, W. J., Kominski, G. F., Ortega, A. N., and Needleman, J. (2010). The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. *Medical care*, 48(7):635–44.

Cifuentes, J., Bronstein, J., Phibbs, C. S., Phibbs, R. H., Schmitt, S. K., and Carlo, W. a. (2002). Mortality in Low Birth Weight Infants According to Level of Neonatal Care at Hospital of Birth. *Pediatrics*, 109(5):745–751.

Costeloe, K., Hennessy, E., Gibson, A. T., Marlow, N., and Wilkinson, A. R. (2000). The EPICure study: outcomes to discharge from hospital for infants born at the threshold of viability. *Pediatrics*, 106(4):659–71.

Department of Health (2003). National Strategy for Improvement. Technical Report April, Department of Health, London.

Draper, E. S., Manktelow, B., Field, D. J., and James, D. (2003). Tables for predicting survival for preterm births are updated. *BMJ (Clinical research ed.)*, 327(7419):872.

Evans, W. N. and Garthwaite, C. (2012). Estimating Heterogeneity in the Benefits of Medical Treatment Intensity. *Review of Economics and Statistics*, 94(3):635–649.

Farley, D. E. and Ozminkowski, R. J. (1992). Volume-outcome relationships and in-hospital mortality: the effect of changes in volume over time. *Medical care*, 30(1):77–94.

Gaynor, M., Seider, H., and Vogt, W. B. (2005). The Volume-Outcome Effect, Scale Economies, and Learning-by-Doing. *The American Economic Review*, 95(2):243–247 CR – Copyright © 2005 American Econom.

Halm, E. a., Lee, C., and Chassin, M. R. (2002). Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Annals of internal medicine*, 137(6):511–20.

Hamilton, K. E. S. C., Redshaw, M. E., and Tarnow-Mordi, W. (2007). Nurse staffing in relation to risk-adjusted mortality in neonatal care. *Archives of disease in childhood. Fetal and neonatal edition*, 92(2):F99–F103.

Luft, H. S., Bunker, J. P., and Enthoven, A. C. (1979). Should operations be regionalized? The empirical relation between surgical volume and mortality. *The New England journal of medicine*, 301(25):1364–9.

Luft, H. S., Hunt, S. S., and Maerki, S. C. (1987). The volume-outcome relationship: practice-makes-perfect or selective-referral patterns? *Health services research*, 22(2):157–82.

McIntire, D. D. and Leveno, K. J. (2008). Neonatal mortality and morbidity rates in late preterm births compared with births at term. *Obstetrics and gynecology*, 111(1):35–41.

Office of National Statistics (2011). Birth cohort tables, England and Wales, 2008.

Office of National Statistics (2012). Birth Summary Tables, England and Wales, 2011 (final).

Parry, G., Tucker, J., and Tarnow-Mordi, W. (2003). CRIB II: an update of the clinical risk index for babies score. *Lancet*, 361(9371):1789–91.

Perrin, K. (2008). *Understanding the Essentials of Critical Care Nursing*. Prentice Hall, New Jersey.

Phibbs, C. S. (2012). Managed clinical networks in neonatal care. *BMJ*, 344(apr03 1):e2423–e2423.

Phibbs, C. S., Baker, L. C., Caughey, A. B., Danielsen, B., Schmitt, S. K., and Phibbs, R. H. (2007). Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *The New England journal of medicine*, 356(21):2165–75.

Rogowski, J. a., Horbar, J. D., Staiger, D. O., Kenny, M., Carpenter, J., and Geppert, J. (2004). Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA : the journal of the American Medical Association*, 291(2):202–9.

Royal College of Pediatrics and Child Health (2012). Annual Report 2011. Technical report, Royal College of Pediatrics and Child Health, London.

Tsai, A. C., Votruba, M., Bridges, J. F. P., and Cebul, R. D. (2006). Overcoming bias in estimating the volume-outcome relationship. *Health services research*, 41(1):252–64.

Tucker, J. (2002). Patient volume, staffing, and workload in relation to risk-adjusted outcomes in a random stratified sample of UK neonatal intensive care units: a prospective evaluation. *Lancet*, 359(9301):99–107.

Tucker, J. and McGuire, W. (2004). Epidemiology of preterm birth. *BMJ (Clinical research ed.)*, 329(7467):675–8.

Tyson, J. E., Parikh, N. A., Langer, J., Green, C., and Higgins, R. D. (2008). Intensive care for extreme prematurity–moving beyond gestational age. *The New England journal of medicine*, 358(16):1672–81.