Article:
Mújica-Mota, RE, Landa, P, Pitt, M et al. (2 more authors) (2020) The heterogeneous causal effects of neonatal care: a model of endogenous demand for multiple treatment options based on geographical access to care. Health Economics, 29 (1). pp. 46-60. ISSN 1057-9230

https://doi.org/10.1002/hec.3970

The heterogeneous effects of neonatal care: a model of endogenous demand for multiple treatment options based on geographical access to care

Running title: Heterogeneous effects of neonatal care

Rubén E Mújica-Mota*[1], Paolo Landa[2], Martin Pitt[2], Mike Allen[2], Anne Spencer[2].

*Corresponding Author. Address: Leeds University School of Medicine, Leeds Institute of Heath Sciences, Room 11.61, Worsley Building, Clarendon Way, Leeds, LS2 9NL. Tel. 0113 3430843. Email: r.e.mujica-mota@leeds.ac.uk.

Affiliations: 1. University of Leeds Medical School, Leeds Institute of Health Sciences, Leeds UK. 2. University of Exeter Medical School, Institute of Health Research, Exeter UK

Acknowledgements: The authors wish to thank Antonieta Medina-Lara for insightful comments, Ekaterina Kuznetsova for thorough discussion, and participants at the 2018 winter meeting of the HESG for helpful suggestions, on an earlier draft of this article. Any errors or omissions remain the sole responsibility of the authors.

Funding: This study was supported by an NIHR Health Services and Delivery Research Programme Grant

Conflicts of Interest: The authors declare no conflict of interest.

The heterogeneous effects of neonatal care: a model of endogenous demand for multiple treatment options based on geographical access to care

Neonatal units in the UK are organised into three levels, from highest Neonatal Intensive Care Unit (NICU), to Local Neonatal Unit (LNU) to lowest Special Care Units (SCU). We model the endogenous treatment selection of neonatal care unit of birth to estimate the average and marginal treatment effects of different neonatal designations on infant mortality, length of stay and hospital costs. We use prognostic factors, survival and hospital care use data on all preterm births in England for 2014-2015, supplemented by national reimbursement tariffs and instrumental variables of travel time from a geographic information system. The data were consistent with a model of demand for preterm birth care driven by physical access. In-hospital mortality of infants born before 32 weeks was 8.5% overall, and 1.2 (95% CI: -0.7, 3.2) percentage points lower for live births in hospitals with NICU or SCU compared to those with an LNU according to instrumental variable estimates. We find imprecise differences in average total hospital costs by unit designation, with positive unobserved selection of those with higher unexplained absolute and incremental costs into NICU. Our results suggest a limited scope for improvement in infant mortality by increasing in-utero transfers based on unit designation alone.

Keywords: Endogeneity, Instrumental variables, control function, multiple treatments, geographical access, semi-parametric, average treatment effects, neonatal, seemingly unrelated regression equations; latent factor, policy evaluation

1 | INTRODUCTION

Preterm birth is accompanied by high risks of morbidity and neonatal mortality, and need for specialised neonatal care services. Since 2003 neonatal services in England are organised into managed clinical networks (DH 2003; Marlow et al. 2007) in which specialist care is centralised and low-level care is distributed across the network. These services are provided in neonatal units of three designated levels of specialisation: level 1 or Special Care units (SCU) look after infants needing level 1 care involving continuous monitoring of their breathing or heart rate, oxygen supply, tube feeding and recovery from phototherapy; level 2 or local neonatal units (LNU) can provide level 1 care as well as providing level 2 care such as short-term intensive care and support including continuous positive airway pressure (CPAP); level 3 or neonatal intensive care units (NICU) can provide level 1 and level 2 care and can additionally provide level 3 care for infants requiring ventilation, CPAP, and weighing <1kg. According to clinical guidelines, births of <28 weeks of gestational age (extremely preterm) should be cared for at a level 3 neonatal unit (NICE 2010). Nevertheless some extremely preterm infants are still born in hospitals with lower level units. Thus the relative effectiveness between neonatal unit designation levels is a key policy issue.

Estimating the relative effects of different designations on infant mortality requires inferring causality from observational data. Infant assignment to hospital of birth may be non-random, thereby confounding the observed mortality differences for true causal effects. Mothers of high-risk preterm infants may seek giving birth at designated level 3 units even among babies of the same gestational age and birthweight (Marlow et al. 2014). Instrument variables (IV) estimation is a method commonly used in economics to infer causality in observational studies (Wooldridge 2010) and increasingly used to estimate causal treatment effects in health service research (Garabedian et al. 2014).

Studies exploring the effects of neonatal unit designation at hospital of birth have shown that low-designated units are associated with increased rates of in-hospital mortality (Lasswell et al. 2010, Phibbs et al. 2007), although a recent study of very low-birthweight infants in California found no such association (Jensen and Lorch 2016). However, differences in organisation structure between UK neonatal services and other nationally funded neonatal services (Kelly et al. 2017), and the much larger neonatal units typical of the US, may limit the generalisability of results across countries. In the UK, Watson et al. estimated the causal effect of level 3 unit on infants born at ≤32 weeks using an instrumental variable (IV) approach and found no evidence that birth in NICU affects in-hospital mortality compared to lower unit designations (Watson et al. 2014a). They also found that higher nurse-to-patient ratios and higher per diem costs reduced infant mortality in NICUs (Watson et al. 2014b, Watson et al. 2017). However, none of these studies sought to analyse unobserved heterogeneity in treatment effects (Cornelissen et al. 2016).

The paper's methodological contribution is to develop an IV estimation framework with the first stage endogenous treatment choice modelled as a demand system, thus providing structural validation tests of identification with continuous geographical access, travel time, IVs. We test for unobserved heterogeneity in marginal treatment effects of NICU vs. other designations combined, and introduce a control function approach for estimating heterogeneous treatment effects with more than two treatment options. These methods are used to estimate the causal effects of preterm birth in a hospital with a SCU, LNU, or NICU, on in-hospital mortality, length of stay and hospital costs.

2 | Causal estimation approach

In this study our IV identification strategy is based on variation in travel time as measure of physical access to treatment. A systematic review of 187 comparative health effectiveness studies using an IV approach between 1993-2011 found that 65 studies had estimated mortality effects and, of these, 27 studies used travel distance (defined as straight line, Euclidean distance, or travel time) as instrumental variable, the second most common instrument after variation in regional treatment patterns (Garabedian et al. 2014).

In our context, IV estimation assumes that study subjects are a mix of high or low risk mothers that by chance live close to a particular unit. The IV estimates based on travel time or distance apply to mothers whose hospital designation at delivery is determined by the relative closeness of different hospital designations, and these mothers are known as 'compliers', because their randomly allocated 'treatment' (i.e. closest unit level) determines their place of delivery. Travel distance or time is a natural predictor of place of birth, and therefore candidate for instrument, as women prefer to deliver in a local unit (Hollowell et al 2016) and birthing units recommend avoiding excessive distances to limit the risk of out-of-hospital birth (Blondel, 2011). Previous distance-based IV studies have used differenced and absolute measures of distance or travel time as instruments in almost equal measure (Garabedian et al. 2014). In this study we use absolute travel times as the more accurate and less restrictive option for a set of IVs and validate them by comparing their actual and expected effects when interpreted as implicit access prices in a model of demand for treatment.

In addition, the continuous scale of both travel time and distance permits us to analyse how treatment effects vary across individuals with different unobserved propensities to use treatments, by estimating marginal treatment effects (MTE, Carneriro, Heckman and Vytlacil 2010), the continuous version of the 'local average treatment effect' (Imbens and Angrist 1994; Angrist and Pishcke 2001). Few studies in health economics have analysed treatment effect heterogeneity (Basu et al. 2007; Basu et al. 2014; Evans and Garthwaite 2012; Tyler-Brown et al. 2011) and this is an aspect we seek to address in this study.

Finally, the IV estimator implies a testable relationship between distance or travel time instruments and demand for the different treatment options. For example, Cutler evaluated heart services using difference in access (distance to hospital of each type) to intervention and control treatments as instruments (Cutler 2007). Watson similarly relied on IV estimation but only used information on the closest hospital and thus ignored instruments on alternative treatment options (i.e. when the closest unit was a non-NICU the characteristics of NICU were omitted and vice versa; Watson et al. 2014). We add to the literature by introducing a control function approach to extend the endogenous heterogeneous treatment effects model to ≥3 treatments.

3 | Methods

3.1 |Data

Data from the National Neonatal Research Database (NNRD) for years 2014 and 2015 were employed in the analysis. The NNRD contains selected information from the BadgerNet Neonatal Electronic Patient Record (https://www.clevermed.com/badgernet/badgernet-neonatal/) on all admissions to NHS neonatal units. Outcomes considered were any in-hospital mortality in the period from birth up to hospital discharge home or to a ward. Data available from the database include antenatal, delivery and neonatal treatments and outcomes. Neonatal unit level designation was taken from the 2015 National Neonatal Audit Programme report (RCPCH 2015). In our sample, 90%

of the 161 neonatal units in England gave permission to access NNRD data (100% of NICUs, 85% of LNUs, and 90% of SCUs). .

Expected fastest road travel times were calculated from a Geographic Information System (Maptitude® 2016) with MPMileCharter® add-in based on coordinates of postcode closest to the population-weighted centroids of the 2011 LSOA (there is one LSOA for each postcode in England) of the parents' residences and closest hospitals of each type and information on typical duration of journey on actual road grid. This created three IVs, i.e. three travel times for each individual, one per neonatal unit level. The 2015 Multiple Index of Deprivation (IMD) for each LSOA was obtained from the Office of National Statistics (ONS 2015). Ethical approval was obtained from the Neonatal Data Analysis Unit at Imperial College, London.

3.2 | Main outcome equation

Three types of infant outcomes are separately analysed: in-hospital mortality, length of hospital stay and associated reimbursement costs, and number of hospital days spent by the infant at three levels of critical care. The binary (mortality), continuous (costs) and discrete count (hospital days) scales of these outcomes required analysis using generalised linear models (Debb and Trivedi 2006) of individual infant outcomes as a function of place of birth (LNU and SCU) relative to a reference unit type (NICU),

$$Y_i = g(\beta_1 SCU_i + \beta_2 LNU_i + X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi}) + v_i \quad (1)$$

where Y_i is an observed continuous or discrete outcome of infant i in the follow-up period up to hospital discharge, SCU_i is a binary variable equal to 1 if the neonatal unit of birth of infant i is SCU and 0 otherwise, and LNU_i is likewise defined for birth in LNU. The term $X_i'\delta$ stands for a linear vector of adjusting covariates commonly used in this literature (Gale et al. 2013, Cole et al. 2010, Manktelow et al. 2013, Ge et al. 2013, Tucker et al. 2002, Lorch et al. 2012; Appendix 0) including birthweight, gestational age, index of multiple deprivation, and number of pregnancies (plus a constant), with their respective coefficients δ.

The terms l_{SCi} and l_{LNi} are unobserved latent utility factors (section 3.3) for SCU and LNU, respectively, that serve to control for the endogeneity of SCU and LNU in Eq. 1, which occurs when coefficients $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$. They account for possible unmeasured confounders, including prognostic factors e.g. congenital abnormalities that place infants at higher risks of neonatal adverse events including death. If, for example, women with high-risk pregnancies choose or are somehow determined by unmeasured factors to deliver at NICUs, a ('naïve') model excluding l_{SCi} and l_{LNi} will incorrectly attribute some of the systematic variation in outcomes to the SCU and LNU variables and likely result in biased estimates of β_{1i} and β_{2i}.

Assuming a mean zero error, $Ev_i = 0$,

$$g^{-1}(EY_i) = \beta_1 SCU_i + \beta_2 LNU_i + X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi} \quad (2)$$

with $g^{-1}(EY)$ denoting the link function (logit or probit for mortality, log for costs, and log for days in hospital) evaluated at the mean of outcome Y, i.e., mortality status, costs, or days in hospital. Eq. 1 is estimated by maximum simulated likelihood, given a suitably chosen parametric distribution for v (binomial for mortality, normal for costs and negative binomial for days in hospital). We estimated

Eq. 1 using IV and control function methods, which required estimating a treatment choice model of the endogenous SCU and LNU binary variables as explained next.

3.3 | Instrumental variables

We evaluate the causal effects on infant outcomes of birth at LNU and SCU vs. NICU hospitals using the three available instruments of travel time to the closest hospital of each neonatal unit type, z_{SC}, z_{LN}, z_{IC}, for SCU, LNU and NICU respectively. At least two instruments were required for estimating treatment effects of the potentially endogenous SCU and LNU variables in Eq. (1) (rank condition; Wooldridge 2010). To be valid, the IVs have to determine the probability of delivering at a LNU and SCU (relevance condition), and be correlated with the outcome Y only through their association with LNU and SCU (conditional independence condition; Appendix 1 eq. 2).The individual may be born in one of three types of neonatal unit, a discrete treatment selection process which we analyse as a multinomial latent demand model where the neonatal unit level in the hospital of birth is the treatment option of maximum latent utility for the mother.

3.4 Demand for hospital type for a very preterm birth

In order to model the endogenous multinomial treatment selection, we define ICU*, LNU*, SCU* as the corresponding latent utilities of birth at the three neonatal unit levels:

$$SCU_i^* = \theta_3 z_{SCi} + \theta_2 z_{LNi} + \theta_1 z_{ICi} + X_i'\gamma_{SC} + \epsilon_{SCi} \quad (3)$$

$$LNU_i^* = \alpha_3 z_{SCi} + \alpha_2 z_{LNi} + \alpha_1 z_{ICi} + X_i'\gamma_{LN} + \epsilon_{LNi}$$

$$ICU_i^* = \pi_3 z_{SCi} + \pi_2 z_{LNi} + \pi_1 z_{ICi} + X_i'\gamma_{IC} + \epsilon_{ICi}$$

where

$$\epsilon_{ji} = W_i'\omega_j + \upsilon_{ji} \quad j = \{SC, LC, IC\}$$

are linear indices of unmeasured demand attributes (W) that are prognostic factors in outcome equation (2) plus an independently distributed random error (υ), while other Greek symbols are coefficients to be estimated. Birth occurs in the unit type of maximum utility:

$$SCU_i = 1 \; if \; SCU_i^* > ICU_i^* \; and \; SCU_i^* > LNU_i^*,$$

$$SCU_i = 0 \qquad otherwise;$$

$$LNU_i = 1 \; if \; LNU_i^* > ICU_i^* \; and \; LNU_i^* > SCU_i^*,$$

$$LNU_i = 0 \qquad otherwise;$$

birth in a NICU occurs when SCU=0 and LNU=0.

We expect $\theta_3<0$, $\alpha_2<0$, and $\pi_1<0$, whilst the coefficients of remaining instrumental variables are expected to be positive or zero. The coefficients of the multinomial choice model of Eq. 3 are not identifiable (Train 2003, p. 26-27). Subtracting the utility of a reference option, say, ICU_i^* from each equation in Eq. 3, results in an identifiable system of two independent equations of differenced utility for SCU and LNU relative to the utility of NICU:

$$\widetilde{SCU}_i^* = \tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi} + X_i' \tilde{\gamma}_{SC} + \tilde{\epsilon}_{SCi} \quad (4a),$$

$$\widetilde{LNU}_i^* = \tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi} + X_i' \tilde{\gamma}_{LN} + \tilde{\epsilon}_{LNi} \quad (4b)$$

where utility differences depend on the three instruments, one for travel time to the closest unit of each type, and the Greek symbols denote the estimable coefficients. The accents denote coefficients transformed by subtracting the corresponding coefficient in the NICU latent equation, and $\tilde{\epsilon}_{LNi}$ and $\tilde{\epsilon}_{SCi}$ are the error terms in the propensity equations after subtracting the error in the NICU latent equation. We expect the own-'access price' effect to be negative ($\tilde{\alpha}_2$<0 and $\tilde{\theta}_3$<0), and the cross-price of access to NICU effect to be positive ($\tilde{\alpha}_1 >$0 and $\tilde{\theta}_1 > 0$). In contrast, the expected signs of $\tilde{\alpha}_3$ and $\tilde{\theta}_2$ are ambiguous a priori (Appendix 1). Birth in NICU (ICU=1) occurs when $\widetilde{SCU}_i^*$<0 in 4a and $\widetilde{LNU}_i^*$<0 in 4b, otherwise, birth occurs in a lower level unit (ICU=0). The case of birth at LNU (LNU=1) and SCU (SCU=1) are defined analogously.

Our control function approach for estimating Eq. 2 (Debb and Trivedi 2006), uses equations 4a & 4b and,

$$\tilde{\epsilon}_{LNi} \equiv \epsilon_{LNi} - \epsilon_{ICi} = W'(\omega_{LN} - \omega_{IC}) + \upsilon_{LNi} - \upsilon_{ICi} \equiv l_{LNi} + \tilde{\upsilon}_{LNi}$$

$$\tilde{\epsilon}_{SCi} \equiv \epsilon_{SCi} - \epsilon_{ICi} = W'(\omega_{SC} - \omega_{IC}) + \upsilon_{SCi} - \upsilon_{ICi} \equiv l_{SCi} + \tilde{\upsilon}_{SCi} \quad (5)$$

where l_{LNi} and l_{SCi} are the values of unobserved indirect utility factors affecting the neonatal outcome Y in Eq. 2. We assume that these terms are distributed standard normal across mothers, and integrate them out of the likelihood function using simulation methods. To derive the likelihood we assume that $\tilde{\upsilon}_{LNi}$ and $\tilde{\upsilon}_{SCi}$ are independently identically extreme-value distributed error terms that are independent from l_{LNi} and l_{SCi} and whose joint distribution implies a multinomial logit treatment choice probability function of the linear indices of covariates and unobserved factors in 4a & 4b (Appendix 2).

In addition, we estimate the multinomial probit treatment choice model (Roodman 2011) that relaxes the independence of irrelevant alternatives (IIA) assumption of the multinomial logit model by allowing the indirect utility equations 4a and 4b to be correlated (Train 2003). In sensitivity analysis we impose the exclusion restrictions on 4a and 4b that all instrument coefficients other than θ_3, α_2, and π_1 equal zero, i.e. $\widetilde{SCU}_i^* = \theta_3 z_{SCi} - \pi_1 z_{ICi} + X_i' \tilde{\gamma}_{SC} + \tilde{\epsilon}_{SCi}$, $\widetilde{LNU}_i^* = \alpha_2 z_{LNi} - \pi_1 z_{ICi} + X_i' \tilde{\gamma}_{LN} + \tilde{\epsilon}_{LNi}$, to address possible issues of identification with this model (Keane 1992; Appendix 1).

3.5 |In-hospital mortality

The endogenous treatment model was specified as a logit outcome with multinomial logit treatment control function (Debb and Trivedi 2006) and, alternatively, as a probit outcome with multinomial probit treatment (Roodman 2011; Appendix 2). We present results in terms of marginal effects.

3.6 |Costs and length of stay

Reimbursement cost and length of hospital stay were analysed as linear outcomes with endogenous multinomial logit (Debb and Trivedi 2006) or probit treatment (Roodman 2011). Reimbursement costs were calculated by multiplying the number of days at each level of care (section 3.7) by the corresponding English 2015 per diem (HRG) tariff. We also estimated heterogeneous treatment effects in correlated random coefficients models (Card 2001), by limited information maximum likelihood (Aakvik, Heckman and Vytalacil 2005; Appendix 2).

3.7 | Inpatient days by level of care

We estimated the effect of neonatal unit designation on the number of days at British Association of Perinatal and Maternity (BAPM) levels of care 1, 2 and 3 separately, which together accounted for 98% of total LOS (the 'super spell' including any post-natal transfer) in our sample. This analysis used a negative binomial endogenous multinomial logit treatment model. We present treatment effect estimates in terms of incidence rate ratios and marginal effects (Appendix 2).

We estimate the MTE of NICU vs. non-NICU birth (Carneiro, Heckman and Vitlacil 2011; Cornelissen et al. 2016) on mortality and the logarithm of hospital costs using a linear endogenous binary treatment model. These analyses use a Gaussian family with an identity link, i.e. a linear probability model for mortality and log linear model for costs. The treatment indicators SCU and LNU in (1) are replaced by a treatment indicator, ICU, equal to 1 when SCU=0 and LNU=0 and 0 otherwise. Also, the strong assumption that the latent factors enter linearly in (1) is relaxed by replacing them with a non-parametric function $K_Y(p)$ of the 'resistance to NICU' treatment or propensity score (p):

$$EY_i = X_i'\delta_{Y0} + X_i'(\delta_{Y1} - \delta_{Y0})p_i + K_Y(p_i) \qquad (6)$$

The MTE is the derivative of (6) with respect to p,

$$MTE_i \equiv \frac{\partial EY_i}{\partial p} = X_i'(\delta_{Y1} - \delta_{Y0}) + \partial K_Y(p)/\partial p$$

MTEs are estimated semi-parametrically (Brave and Walstrum 2014) and plotted relative to p. We estimate alternative MTEs under the parametric probit treatment choice model (Appendix 3).

We tested for the existence of unobserved selection by prognosis (H0: $\rho_1 = 0$), where infants who have worse unobserved prognosis may be more likely to be born in NICU than infants with better prognosis, and selection by returns (H0: $\partial K_Y(p)/\partial p = 0$ in (6) or $\sigma_1\rho_1 - \sigma_0\rho_0 = 0$ in (7)), where infants with unobserved characteristics predisposing them to benefit more from treatment are more likely to be born in NICU (Appendix 3).

Standard errors are calculated using the method by White (1980), to account for clustering of infants in hospitals, except for MTEs, which are estimated at the mean of covariates X, using the bootstrap percentile method. Stata code illustrating the implementation of main analyses is provided in Appendix 4.

4 | RESULTS

4.1 | Distribution of sample characteristics by geographical access

Data on 14,727 live births at less than 32 weeks' gestation were available from the NNRD, 12,990 of which had complete data on infant and hospital characteristics for analysis, with 303 observations having invalid data values. Of the 12,687 remaining observations, 1650 (13%) individuals had no travel time to the closest SCU or LNU hospitals data and were excluded from the analysis. The remaining sample included 11,037 patients from 154 hospitals, of which 11 were hospitals that delivered at least 100 infants weighing <1500 g per year on average during the study period ('high-volume'); all of these hospitals were ICU and 42% (2377) of the 5595 infants born in a NICU level were delivered in a high-volume hospital. Fifteen infants were born in a hospital without a neonatal unit and were transferred ex-utero to the closest neonatal unit in the network (14 to SCU, 1 to LNU);

they were analysed according to the level of these units. In-hospital mortality in the analysis sample was 8.52% (8.43% including missing travel time data cases).

There are no systematic differences in most descriptive characteristics of the analysed sample across travel time to NICU tertiles (Table 1). In addition to the exposure variables (delivery at NICU, LNU and SCU), systematic differences arise only for deprivation of residence, unknown delivery mode and suggest the need to control for possible confounding by these variables in our analyses. Similar results were obtained for tables in terms of travel times to LNU and SCUs and in London (Appendix 5).

4.2 | Demand (choice) model–first stage

Table 2 presents estimates obtained from multinomial probit and multinomial logit models for the probability of birth in LNU (second and fourth columns) and the probability of birth at SCU (third and fifth columns), adjusted for covariates. The signs of these coefficients are consistent with our a priori expectations. The two coefficients with ambiguous expectations a priori, the cross-price effects of access to SCU in the LNU equation and to LNU in the SCU equation are negative with $p>0.10$, suggesting that the effect of travel time to LNU on the utility of SCU, and vice versa, is equal to or smaller than its effect on the utility of NICU (eq. 4a, and $b). The probability of birth in a LNU level facility was positively related with longer travel times to the closest NICU, and with longer travel times to the closest SCU, whereas being negatively related with longer travel times to the closest LNU facility. The price elasticity of demand decreases with level of specialisation, with NICU care being the least responsive option to an increase in its own travel-time price of access. Birth at SCU is nine times as responsive to travel time to NICU as it is to travel time to LNU (0.61 vs. 0.07).

4.3 | Estimates of in-hospital mortality

Table 3 summarises the estimated marginal effects of birth at LNU vs NICU and birth at SCU vs. NICU in the naïve single equation probit model (second column) and corresponding average treatment effects of the IV model that adjusts for unobserved confounding (third column). In the naïve probit model birth in a SCU is associated with a 1.7 percentage point higher risk of neonatal death than birth in NICU ($p=0.09$), while LNU with 0.4 percentage point excess risk over NICU ($p=0.54$). In the IV model, the respective estimates are 0.1 ($p=0.96$) and 1.2 ($p=0.23$) under a probit specification. According to the IV model diagnostic statistics, the hypothesis that birth at SCU is exogenous cannot be rejected at $p=0.05$. Results were similar for logit specifications.

Similar results were obtained in the subgroup of infants born at less than 28 weeks' gestation (Appendix 5).

Our main results (reproduced in Table 4 column a) were robust to excluding socio-economic and including mode of delivery covariates, and to variation in the specification of the endogenous treatment model. Moreover, tests on the estimated correlations between the random error terms of

the multinomial treatment equations and the mortality equation do not reject the null hypothesis that birth at LNU and birth at SCU are exogenous in the mortality equation at p=0.05 under both logit and probit specifications.

The IV estimates will not apply to those mothers who deliver in NICUs regardless of the distance or time required to travel from home to their closest NICU. For example, high-risk mothers with history of preterm birth may be booked in for birth at a hospital with a NICU in spite of it not being their closest hospital; the so-called *always takers* of the intervention (birth at NICU) regardless of travel time. The IV estimates will also not apply to high-risk mothers who are not transferred to higher level units because of their infants' poor life prospects; the so-called *never takers* of birth at NICU. The proportion of *always takers* in our dataset appears to be higher than the proportion of *never takers*: 732 (22%) of those mothers who would need more time to reach their closest NICU than to reach their closest LNU and their closest SCU would still deliver at a NICU; in contrast, only 56 (1.5%) and 388 (8.9%) mothers whose closest (minimum travel time) hospital was a NICU delivered in a SCU and LNU, respectively. The analysis of MTE of birth at hospitals with NICUs vs. hospitals with a lower-designation neonatal unit produced treatment effect estimates with 95% CI crossing zero throughout the unobserved resistance to NICU treatment (Appendix 6).

4.4 | Estimates on length of stay and costs

The estimated total duration of the neonatal hospital stay including hospital transfers (i.e. the 'super spell') of an infant born in NICU, LNU and SCU was, respectively 66, 66, and 67 days (differences: SCU vs NICU 1.0, p=0.76; LNU vs. NICU 0.6, p=0.81; Appendix 7 Table A7.1). The reimbursement cost of birth was respectively £42,776, £44,854 and £43,220 per infant (NICU minus LNU, -£2078 [95% CI: -5551,1396]; NICU minus SCU, -£444 [-4690,3802]). The results for reimbursement cost and LOS (Appendix 7 Table A7.1) are robust to varying the covariates (available from the authors).

Different test results for homogeneous effects were obtained for LNU ($p<0.05$) and SCU ($p>0.05$) using a control function approach. Unobserved characteristics that led mothers to prefer LNU over ICU were also associated with lower in-hospital costs; e.g. conditional on covariates, mothers in the top 16 percent LNU utility ranking cost under £4634 less than the average. Moreover, individuals with below-average unobserved LNU utility factors (i.e. ceteris paribus above-average NICU utility, eq. 4b) have above-average returns (cost savings vs. NICU) with LNU (Appendix 6).

Parametric normal MTE for NICU vs non-NICU had 95% CI that crossed zero (H0: no positive selection into NICU by non-observably more costly patients, p=0.001; more incrementally costly patients, p=0.17) (Appendix 7). Semi-parametric analysis reveals, however, that mothers who delivered in NICU despite having the 20 to 40 percentile lowest predicted probabilities of doing so ('unobserved resistance' on the x-axes in Figure 1) have the highest incremental costs relative to a non-NICU birthplace.

While birth at lower level units results in very preterm infants spending the same total number of days in hospital as they would if born at a NICU, birth at LNU results in more intensive care (BAPM 1) days (IRR 1.40, 95% CI: 1.26,1.55) and fewer specialised intensive care (BAPM 3) days (IRR 0.95, 95% CI: 0.90,1.01) relative to what would happen if the same infant were born in ICU (or SCU; Figure 2). Birth at SCU results in similar numbers of inpatient days of treatment at the three levels of care relative to birth at NICU (Appendix 9).

5 | DISCUSSION

Our study found that the occurrence of very preterm births outside NICUs was consistent with a model of demand for preterm birth care driven by physical access. Using data on physical access as instrumental variables produced a 0.9-1.3 percentage points lower mortality in NICU and SCU relative to LNU. In contrast, in the simple naïve model with common prognostic covariates, in-hospital mortality was 1-2 percentage points lower in hospitals with NICU or LNU compared to those with an SCU. The 95% CI of all these estimated differences crosses zero, suggesting they are due to chance alone.

We found that our data were compatible with a mortality model in which there is no unobserved confounding. In cases without such confounding, the IV method is inefficient relative to simple regression analysis and may lead to incorrect inferences (Wooldridge 2010). However, we have a priori reasons to suspect endogeneity is present e.g. from selective choice of NICU by pregnancies with risk factors not recorded in our data, and our instruments were found to be strong and valid. Therefore the likely treatment effect for designated units lies with the IV results. Moreover, there is no evidence of an increase in the total length of the infant stay in these neonatal units or cost to commissioners when these outcomes are analysed unadjusted for the competing death risk. Since there are few 'never NICU takers', our IV estimates may be interpreted as the treatment effect on the NICU-untreated (Angrist and Pischke 2009). Thus our results suggest that increases of in-utero transfers from lower unit designations alone are unlikely to bring large improvements in in-hospital mortality (Gale et al. 2012a,b).

Our study also exploited continuous instruments to analyse the heterogeneity in treatment effects on mortality and costs. Our results failed to reject the hypothesis that there is no residual unobservable self-selection of women into NICU according to neonate severity or expected mortality risk reduction at conventional significance levels; however, it is possible that a larger sample would have rejected it. In terms of costs, there is evidence of unobservable self-selection of complex (i.e. more costly) cases into NICU hospitals and of negative selection by returns as some infants with the highest additional costs relative to non-NICU care are prone to be born in NICU hospital for reasons unrelated to birthweight, gestational age, socio-economic status, number of pregnancies and sex.

We found a significant causal reduction in the number of hospital days spent under the most intensive care level (BAPM 1) that was accompanied by an increase in the number of days under lower care intensity (BAPM 3) with NICU relative to LNU. While the associated net effect on overall reimbursement costs to the NHS is apparently zero, and we did not find the mortality benefits documented by Marlow and colleagues (Marlow et al. 2014), these results suggest nevertheless that birth at NICU would reduce neonatal morbidity among those currently born in LNU. Further research that investigates this question is warranted using measures of neonatal morbidity including ventilator days; bronchopulmonary dysplasia; intraventricular haemorrhage, particularly the severe grades 3-4; late-onset infection; necrotizing enterocolitis; and retinopathy of prematurity, particularly severe stages 3 and above.

A limitation of our analysis is that the IV method requires the assumption that travel time to the closest neonatal unit did not affect infant mortality by means other than through its role in determining the level of the neonatal unit of the hospital of birth. It is possible that longer travel time to a NICU increased the chance of in-hospital mortality among those infants delivered in a NICU due to delays in receiving the required specialised care. However, we would expect these effects, if present, to be secondary to the effects of travel time on mortality that are due to exposure to the level of care of the neonatal unit of birth.

Our measure of mortality, in-hospital infant death, did not include stillbirths, which exceed neonatal deaths in England (2952 versus 1721 annually, ONS 2015). Another limitation of our dataset is its lacking information on antenatal steroid use (ANS), which may account for the poorer mortality results for the SCUs as these use less steroids (RCPCH 2017). Watson et al. using the same database reported that covariates, including ANS, were evenly distributed between NICU and non-NICU born very preterm infant groups, after controlling for the lowest decile of index of multiple deprivation (Watson et al. 2015). We thus expect any omitted variable bias from ANS in our analysis, after controlled for quintiles of socioeconomic deprivation, to be limited. Low socio-economic status is itself linked to an increased risk of preterm births through low maternal weight and smoking (Taylor-Robinson et al. 2011). Therefore, any unmeasured differences in socio-economic status that are not captured by our multiple deprivation measure may have confounded our results also.

Future work should investigate differences in mortality and costs between high and low-volume NICUs since a high volume of births may be more influential on neonatal mortality and outcomes than a high designation level of unit (Jensen and Lorch 2015). Our findings comprise 42% of NICU infants born in high-volume units in our sample.

Supplementary files

Appendix 0 – Covariates included in the analysis

Appendix 1 – Identification of the instrumental variable estimator

Appendix 2 – Main outcome equation models

Appendix 3 – Marginal Treatment Effect analysis

Appendix 4 – Stata code to implement the main analyses

Appendix 5 – Distribution of sample characteristics by travel time to SCU and LNU

Appendix 6 – Main results for extremely preterm births

Appendix 7 – Results on LOS and reimbursement costs for VPT births

Appendix 8 – Results on inpatient days by level of care for VPT births

Appendix 9 – Results on marginal treatment effects

References

1. Aakvik, A., Heckman, J.J. and Vytlacil, E.J., 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics*, *125*(1-2), pp.15-51.
2. Angrist, J. and J. Pischke, *Mostly harmeless Econometrics An empiricist's companion*. 2009, Princeton, New Jersey, USA: Princeton University Press. 373.
3. Basu A, Heckman J, Navarro-Lozano S, Urzua S. 2007. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* 16(11): 1133–1157.
4. Basu A, Jena A, Goldman DP, Philipson TJ, Diubois R. Heteroegenity in action: the role of passive personalization in comparative effectiveness research. Health Economics 2014; 23: 359-373.
5. Blondel B, Drewniak N, Pilkington H, Zeitlin J. Out-of-hospital births and the supply of maternity units in France. Health Place 2011;17:1170–3. https://doi.org/10.1016/j.healthplace.2011.06.002
6. Brave S,Walstrum T. Estimating marginal treatment effects using parametric and semiparametric methods. Stata J. 2014. 14 (1), 191–217
7. Brown, T. T., Dela Cruz, E. and Brown, S. S. (2011), The effect of dental care on cardiovascular disease outcomes: an application of instrumental variables in the presence of heterogeneity and self-selection. Health Econ., 20: 1241-1256. doi:10.1002/hec.1667
8. Caliper, *Mapping & Transportation Software Solutions*. 2017.
9. Card D. Estimating the return to schooling: progress on some persistent econometric problems. Econometrica 2001; 69 (5), 1127–1160.
10. Carneiro, P., Heckman, J.J., Vytlacil, E.J., 2011. Estimating marginal returns to education. Am. Econ. Rev. 101 (6), 2754–2781.
11. Cole, T.J., E. Hey, and S. Richmond, The PREM score: a graphical tool for predicting survival in very preterm births. Arch Dis Child Fetal Neonatal Ed, 2010. 95(1): p. F14-9.
12. Cornelissen T, Dustmman C, Raute A, Schonberg U. From LATE to MTE: Alternative methods for the evaluation of policy interventions. Labour Economics 2016;41:47-60.
13. Cutler DM. The lifetime costs and benefits of medical technology. Journal of Health Economics, 2007. **27**: 1081-1100.
14. Debb, P. and P. Trivedi, *Specification and simulated likelihood estimation of a non-normal treatment-outcone model with selection: application to healthcare utilization.* The Econometrics Journal, 2006. **9** (2): p. 307–331.
15. *Department of Health, Report of Department of Health Working Group on Neonatal Intensive Care Services, London.* 2003.
16. Evans WN, Garthwaite C. Estimating heterogeneity in the benefits of medical treatment intensity. The Review of Economics and Statistics 2012; 94(3): 635-649.
17. Gale, C., et al., *Impact of managed clinical networks on NHS specialist neonatal services in England: population based study.* BMJ, 2012a. **344**.
18. Gale, W.J., et al., Prediction of Neonatal Outcomes in Extremely Preterm Neonates. Pediatrics, 2013. 132(4): p. E876-E885.
19. Gale C, Hay A, Philipp C, Khan R, Santhakumaran S, Ratnavel N. In-utero transfer is too difficult: Results from a prospective study. Early Human Development 2012b; 88, 147–150.
20. Garabedian LF, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research. Ann Intern Med. 2014;161:131–138. doi: 10.7326/M13-1887.
21. Ge WJ, Mirea L, Yang JM, Bassil KL, Lee SK, Shah PS, et al. Prediction of Neonatal Outcomes in Extremely Preterm Neonates. Pediatrics. 2013;132(4):E876-E85.
22. Greene W. Econometric analysis. 5th edition. Prentice Hall, New Jersey USA, 2003. 1026 pp.
23. Heckman JJ, Urzua S, Vitlacyl E. Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics* 2006 88 (3): 389–432.13.

24. Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences, 1999. 96(8): p. 4730.
25. Hollowell, J., Li, Y., Malouf, R., Buchanan, J.: Women's birth place preferences in the United Kingdom: a systematic review and narrative synthesis of the quantitative literature. BMC Pregnancy Childbirth **16**(1), 213 (2016). doi:10.1186/s12884-016-0998-5
26. Imbens, G.W. and J.D. Angrist, *Identification and Estimation of Local Average Treatment Effects.* Econometrica 1994. **62**(2): p. 467–475.
27. Jensen, E.A. and S.A. Lorch, *Effects of a Birth Hospital's Neonatal Intensive Care Unit Level and Annual Volume of Very Low-Birth-Weight Infant Deliveries on Morbidity and Mortality.* JAMA Pediatrics, 2015. **169**(8): p. e151906.
28. Keane, M., *A note on identification in the multinomial probit model.* Journal of Business and Economics Statistics, 1992. **10**(2): p. 193-200.
29. Kelly, L.E., et al., *Perinatal health services organization for preterm births: a multinational comparison.* Journal of Perinatology, 2017. **37**(7): p. 762-768.
30. L.L.C., W.S.T., *MileCharter: Create Mileage Charts with Maptitude*. 2017.
31. Lasswell SM, Barfield WD, Rochat RW, Blackmon L. Perinatal Regionalization for Very Low-Birth-Weight and Very Preterm Infants A Meta-analysis. Jama-Journal of the American Medical Association. 2010;304(9):992-1000.
32. Lee, H.C., et al., Prediction of death for extremely premature infants in a population-based cohort. Pediatrics, 2010. 126(3): p. e644-50.
333. Lorch, S.A., et al., The Differential Impact of Delivery Hospital on the Outcomes of Premature Infants. Pediatrics, 2012. 130(2): p. 270-278.
34. Manktelow, B.N., et al., Population-Based Estimates of In-Unit Survival for Very Preterm Infants. Pediatrics, 2013. 131(2): p. E425-E432.
35. Marlow, N. and A.B. Gill, *Establishing neonatal networks: the reality.* Archives of Disease in Childhood-Fetal and Neonatal Edition, 2007. **92**(2): p. F137-F142.
36. Marlow, N., et al., *Perinatal outcomes for extremely preterm babies in relation to place of birth in England: the EPICure 2 study. .* Archives of Disease in Childhood - Fetal and Neonatal Edition 2014. **99**.
37. Medlock S, Ravelli AC, Tamminga P, Mol BW, Abu-Hanna A. Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One. 2011;6(9):e23441.
38. NICE, *Specialist neonatal care quality standard*. 2010: London. p. 36.
39. ONS. *English Indices of Deprivation 2015 - LSOA Level 2015*. 2015 [cited 2017; Available from: https://data.gov.uk/dataset/english-indices-of-deprivation-2015-lsoa-level.]
40. ONS. Live births, stillbirths, neonatal deaths and infant deaths, by local authority, age of mother and index of multiple deprivation, England, 2015 [Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/stillbirths/adhocs/]
41. Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. The New England Journal of Medicine. 2007;356:2165-75.
42. RCPCH, *National Neonatal Audit Programme 2017 Annual Report on 2016 data.* . 2017, Royal College of Paediatrics and Child Health.
43. Roodman, D., *Fitting fully observed recursive mixed-process models with cmp.* The Stata Journal, 2011. **11**(2): p. 159-206.
44. Rubin DB. Causal inference using potential outcomes: Design, modelling, decisions.. Journal of the American Statistical Association, 2005. 100(469):322-331.
45. Sivey, P. (2012), The effect of waiting time and distance on hospital choice for English cataract patients. Health Econ., 21: 444–456. doi:10.1002/hec.1720

46. Taylor-Robinson, D., Agarwal, U., Diggle, P.J., Platt, M.J., Yoxall, B., Alfirevic, Z.: Quantifying the Impact of Deprivation on Preterm Births: A Retrospective Cohort Study. Plos One **6**(8) (2011). doi:10.1371/journal.pone.0023163
47. Tucker, J., et al., 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, 2002. 359(9301): 99-107.
48. Watson, S., Arulampalam, W., Petrou, S.: The effect of health care expenditure on patient outcomes: Evidence from English neonatal care. Health Econ. 2017: 1–11.
48. Watson, S., et al., The effects of designation and volume of neonatal care on mortality and morbidity outcomes of very preterm infants in England: retrospective population-based cohort study. BMJ Open 2014. **4**: p. e004856.
50. Watson SI, Arulampalam W, Petrou S, Marlow N, Morgan AS, Draper ES, et al. The effects of a one-to-one nurse-to-patient ratio on the mortality rate in neonatal intensive care: a retrospective, longitudinal, population-based study. Arch Dis Child Fetal Neonatal Ed. 2016;101(3):F195-200.
51. Wooldridge, J., *Econometric analysis of cross section and panel data*. 2010, Cambridge, Massachussetts 02142: Massachussetts Institute of Technology Press. 1064.

Table 1 Sample characteristics by travel time to NICUs (% unless stated otherwise)

	All available observations (N=12,687)			Excluding cases with missing travel time to LNU or SCU data (N=11,037)		
	Lower tertile N=4,191	Medium tertile N=4,185	High tertile N=4,311	Lower tertile N=3,855	Medium tertile N=3,500	High tertile N=3,682
Died	8.26	8.89	8.14	8.40	9.17	8.04
Discharged home	87.30	84.49	81.52	87.16	85.77	86.77
Discharged ward	1.29	1.53	2.07	1.37	1.66	2.01
Last record: transferred to another hospital/unit	2.94	4.56	7.66	2.88	3.20	3.02
Unknown destination	0.21	0.55	0.60	0.18	0.20	0.16
Gestational age at birth (weeks), mean (SD)	28.41 (2.37)	28.43 (2.33)	28.47 (2.30)	28.38 (2.38)	28.46 (2.33)	28.53 (2.29)
Birthweight (kg), mean (SD)	1.19 (0.38)	1.20 (0.38)	1.21 (0.39)	1.18 (0.38)	1.20 (0.38)	1.22 (0.39)
Foetus 2+	25.84	27.60	27.70	26.04	27.49	27.59
Female sex	46.36	45.16	46.69	46.46	44.80	45.95
Residence: Most deprived quintile[1]	47.67	29.49	21.76	49.55	30.36	22.44
Residence: 2nd most deprived quintile[1]	23.00	24.35	21.87	22.65	23.74	21.27
Residence: 3-5 least deprived quintile[3]	29.33	46.16	56.36	27.80	45.90	56.29
Caesarean delivery	48.34	50.75	51.54	48.50	51.14	51.54
Spontaneous vaginal	37.06	36.92	36.67	37.15	36.46	36.85
Unknown delivery mode	4.84	3.70	0.00	4.77	3.74	0
Delivery at NICU	81.53	42.39	26.70	83.24	43.00	23.93
Delivery at LNU	13.36	47.22	58.76	11.47	46.11	60.81
Delivery at SCU	5.11	10.39	14.54	5.29	10.89	15.26
Delivery at high volume[2]	32.38	19.52	10.69	34.42	19.40	10.08

[1] Ranked by the index of multiple deprivation of residential postcode. [2] Defined as born in hospital delivering more than 100 infants with <1500 g birthweight per year during the study period. SD: Standard deviation.

Table 2 Linear index coefficients of instruments in IV multinomial treatment models

	Multinomial probit		Multinomial logit		Elasticities (multinomial logit)		
Instrumental variable	Birth at LNU†	Birth at SCU†	Birth at LNU†	Birth at SCU†	Birth at ICU	Birth at LNU	Birth at SCU
	N= 11,037	N= 11,037	N= 11,037	N= 11,037			
Minimum travel time (mins) to NICU	0.063*** (0.006)	0.075*** (0.024)	0.087*** (0.009)	0.076*** (0.013)	-1.34 (-1.72, -0.97)	0.80 (0.58, 1.02)	0.61 (0.06, 1.16)
Minimum travel time (mins) to LNU	-0.064*** (0.008)	-0.034 (0.021)	-0.109*** (0.013)	-0.026 (0.016)	0.69 (0.47, 0.92)	-1.55 (-1.96, -1.14)	0.07 (-0.49, 0.63)
Minimum travel time to SCU	-0.014 (0.009)	-0.112 (0.105)	-0.003 (0.012)	-0.152*** (0.017)	0.27 (-0.07, 0.60)	0.13 (-0.24, 0.51)	-4.14 (-4.95, -3.33)
Wald F test Ho: all instruments have no effect	188***	36***	153***	172***	N/A	N/A	N/A
Correlation across equations (ρ_{13})	0.82**		Not allowed		N/A	N/A	N/A

†1=yes; 0=no equation. *Controlled covariates: Age and age squared at birth, birthweight, birthweight squared, sex, deprivation of residence, foetus no. N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01.N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01. Statistical inferences based on robust standard errors adjusting for clustering of observations by hospital. Figures in parentheses are standard errors except under elasticities, which are 95% CI.*

Table 3 Causal effects on mortality of birth in LNU & SCU relative to ICU in infants born at <32 weeks

	Naïve	**IV**		**Naïve**	**IV**
	Probit regression	**Probit with endogenous multinomial probit treatment**	**Probit with endogenous multinomial probit treatment – with exclusion restrictions**	**logit regression**	**Logit with endogenous multinomial logit treatment**
Birth at LNU ([0,1] range)	0.004 (0.007)	0.012 (0.010)	0.013 (0.009)	0.006 (0.008)	0.012 (0.010)
Birth at SCU ([0,1] range)	0.017* (0.010)	0.001 (0.015)	-0.001 (0.015)	0.020* (0.010)	0.003 (0.017)
ρ_{12}, λ_1		-0.04	-0.06		-0.202
ρ_{13}, λ_2		0.07	0.12		0.391
ρ_{23}		0.82*	0.20		
Instrument strength: Wald F test statistic (3 degrees of freedom)	N/A	LNU equation: 191*** SCU equation: 34***	LNU equation: 200*** SCU equation: 98***	N/A	LNU equation: 171*** SCU equation: 151***
N	**11,037**	**11,037**	**11,037**	**11,037**	**11,037**
Hausman test z statistic of H0: no endogeneity LNU treatment variable	N/A	-0.63	-1.23	N/A	-1.0
Hausman test z statistic of H0: no endogeneity SCU treatment variable	N/A	0.77	1.32	N/A	1.1
z statistic: no correlation between utility equations (IIA)	N/A	1.82*	0.59	N/A	H0 true by implicit assumption
Test z statistic Ho: valid over-identifying restriction of minimum travel time to NICU	N/A	-0.42	0.26	N/A	0.59

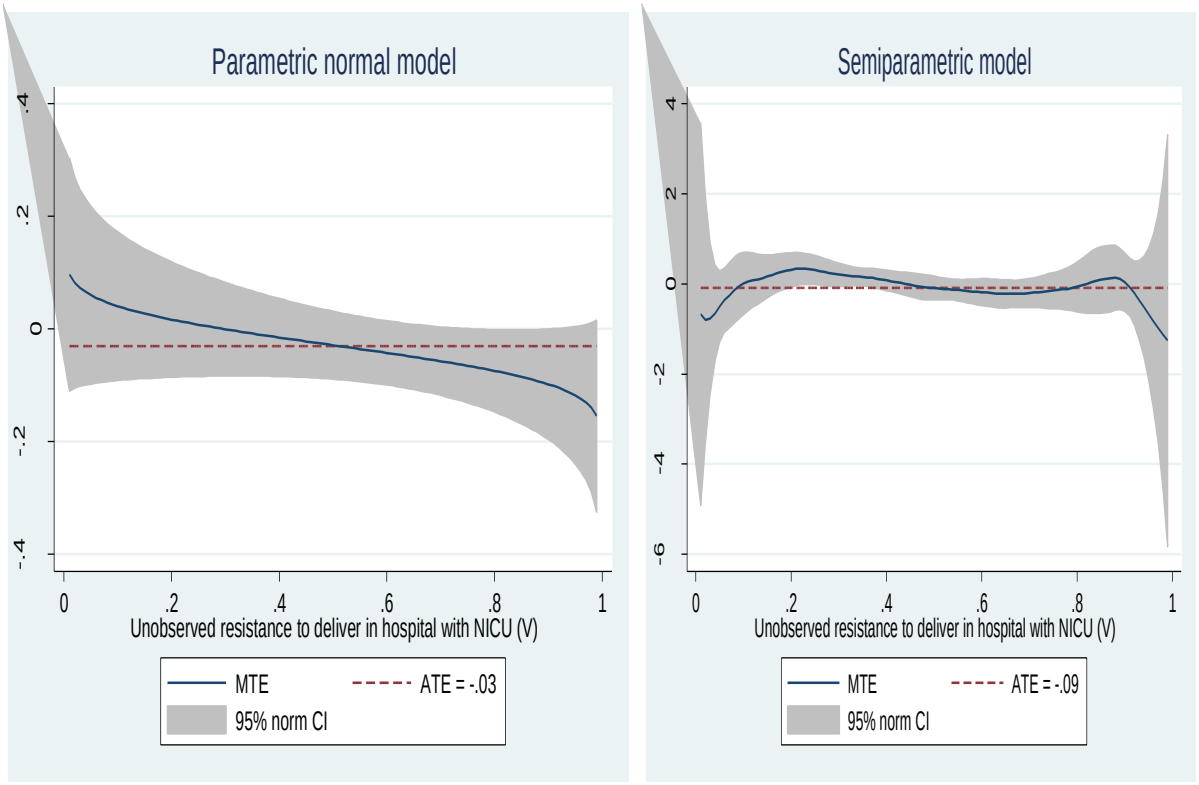
*Controlled covariates: Age and age squared at birth, birthweight, birthweight squared, sex, deprivation of residence, foetus no. N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01.N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01. Statistical inferences based on robust standard errors (in parentheses) adjusting for clustering of observations by hospital.*

Table 4 Robustness check: marginal effects on mortality of birth in LNU & SCU relative to ICU

	Multinomial probit treatment model			**Multinomial logit treatment model**		
	(a)	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**
Birth at LNU (difference [0,1] range)	0.012 (0.010)	0.009 (0.010)	0.011 (0.010)	0.012 (0.010)	0.010 (0.010)	0.011 (0.010)
Birth at SCU (difference [0,1] range)	0.001 (0.015)	-0.003 (0.015)	0.000 0.015)	0.003 (0.017)	-0.000 (0.017)	0.003 (0.017)
Included Covariates?						
Gestational age (GA), GA squared	Yes	Yes	Yes	Yes	Yes	Yes
birthweight, birthweight squared	Yes	Yes	Yes	Yes	Yes	Yes
Infant's sex	Yes	Yes	Yes	Yes	Yes	Yes
Foetus number	Yes	Yes	Yes	Yes	Yes	Yes
Quintiles of multiple deprivation index	Yes	No	Yes	Yes	No	Yes
Mode of delivery and labour	No	No	Yes	No	No	Yes
Instrument strength: Wald F test statistic (3 degrees of freedom)	LNU equation: 191*** SCU equation: 34***	LNU equation: 190*** SCU equation: 38***	LNU equation: 188*** SCU equation: 36***	LNU equation: 171*** SCU equation: 151***	LNU equation: 167*** SCU equation: 146***	LNU equation: 172*** SCU equation: 154***
N	**11,037**	**11,037**	**11,037**	**11,037**	**11,037**	**11,037**
z statistic of H0: no endogeneity LNU treatment	-0.63	-0.38	-0.57	-1.05	-0.82	-1.01
z statistic of H0: no endogeneity SCU treatment	0.77	0.92	0.77	1.06	1.20	1.04

*N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01.N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01. Statistical inferences based on robust standard errors (in parentheses) adjusting for clustering of observations by hospital.*

Figure 1. Marginal Treatment Effects on hospital reimbursement costs (in logarithms) of level 3 vs. lower designation hospital



Notes: Parametric model was estimated as log linear model and a probit model for NICU vs.non-NICU birth under the Potential outcomes framework. The x-axis depicts the unobserved resistance to treatment, V in Appendix 3, which equals the predicted probability of treatment in the first stage choice model. Semiparametric model is the local IV estimator (Heckman and Vytlacil 1999) as implemented by Brave and Walstrum (Brave and Walstrum 2014).

Appendix 0 – Covariates included in the analysis.

The dataset includes common covariates in this literature (Gale et al. 2013, Cole et al. 2010, Manktelow et al. 2013, Ge et al. 2013, Tucker et al. 2002, Lorch et al. 2012): gestational age in weeks, gestational age in weeks squared, birth weight (kg), birth weight squared, index of multiple deprivation (quintiles), number of foetuses (1 vs. 2+), female sex (yes vs. no), mode of delivery and labour (spontaneous vaginal, induced vaginal, emergency caesarean with labour, emergency caesarean without labour, elective caesarean, unknown). We also have information on the mode of delivery to use as control for potential confounding, e.g. since we did not have data available on fetal deaths, to avoid underestimating the health benefits of level 3 units' ability for rapid delivering a preterm birth who may otherwise die in utero (Jensen and Lorch 2015). Nevertheless, we also conducted sensitivity analyses that exclude mode of delivery variables. We considered models using birthweight by gestational age z score instead of birthweight and its squared value but since the two specifications produced the same results we preferred the latter.

Appendix 1 – Identification of the instrumental variable estimator

We Defining ICU*, LNU*, SCU* as the corresponding latent utilities of birth at the three levels of care:

$$SCU_i^* = \theta_3 z_{SCi} + \theta_2 z_{LNi} + \theta_1 z_{ICi} + X_i'\gamma_{SC} + \epsilon_{SCi} \quad (3)$$

$$LNU_i^* = \alpha_3 z_{SCi} + \alpha_2 z_{LNi} + \alpha_1 z_{ICi} + X_i'\gamma_{LN} + \epsilon_{LNi}$$

$$ICU_i^* = \pi_3 z_{SCi} + \pi_2 z_{LNi} + \pi_1 z_{ICi} + X_i'\gamma_{IC} + \epsilon_{ICi}$$

we expect $\theta_3<0$, $\alpha_2<0$, and $\pi_1<0$, whilst the coefficients of the remaining instrumental variables are expected to be positive or zero, e.g. longer travel times to the closest alternative care levels of LNU and ICU are expected to increase the propensity of birth in SCU. The multinomial choice model is estimated by defining LNU* and SCU* as differences in utility relative to the utility of NICU:

$$\widetilde{SCU}_i^* \equiv SCU_i^* - ICU_i^* = (\theta_3 - \pi_3)z_{SCi} + (\theta_2 - \pi_2)z_{LNi} + (\theta_1 - \pi_1)z_{ICi} + X_i'(\gamma_{SC} - \gamma_{IC}) + \epsilon_{SCi} - \epsilon_{ICi} \equiv \tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi} + X_i'\tilde{\gamma}_{SC} + \tilde{\epsilon}_{SCi} \quad (4a),$$

$$\widetilde{LNU}_i^* \equiv LNU_i^* - ICU_i^* = (\alpha_3 - \pi_3)z_{SCi} + (\alpha_2 - \pi_2)z_{LNi} + (\alpha_1 - \pi_1)z_{ICi} + X_i'(\gamma_{LN} - \gamma_{IC}) + \epsilon_{LNi} - \epsilon_{ICi} \equiv \tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi} + X_i'\tilde{\gamma}_{LN} + \tilde{\epsilon}_{LNi} \quad (4b)$$

$$\widetilde{ICU}_i^* \equiv ICU_i^* - ICU_i^* = 0 \quad (4c)$$

where the utility differences depend on the three instruments, one for travel time to the closest unit of each type, and the Greek symbols denote the coefficients of the ICU, LNU or SCU propensity equations, the accents denote coefficients transformed by subtracting the corresponding coefficient in the NICU latent equation, and $\tilde{\epsilon}_{LNi}$ and $\tilde{\epsilon}_{SCi}$ are the error terms in the propensity equations after subtracting the error in the NICU latent equation. We expect the own-'access price' effect to be negative ($\tilde{\alpha}_2<0$ and $\tilde{\theta}_3<0$), and the cross-price of access to NICU effect to be positive ($\tilde{\alpha}_1>0$ and $\tilde{\theta}_1>0$). In contrast, the expected signs of $\tilde{\alpha}_3$ and $\tilde{\theta}_2$ are ambiguous a priori; if $\tilde{\alpha}_3>0$, then $\theta_3 > \pi_3$ and the effect of access to SCU is larger on the utility of LNU than on the utility of ICU, and viceversa if $\tilde{\alpha}_3<0$; the same situation applies to $\tilde{\theta}_2$.

Turning to the neonatal treatment decision criteria in the multinomial model, birth in NICU (ICU=1) occurs when its latent utility is the highest of the three hospital levels, i.e., LNU*< ICU* and SCU*<ICU*, which in terms of 4a and 4b implies, respectively:

$$\tilde{\epsilon}_{SCi} < -\ (\tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi} + X_i' \tilde{\gamma}_{SC}) \qquad \text{(A1.4a)}$$

&

$$\tilde{\epsilon}_{LNi} < -\ (\tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi} + X_i' \tilde{\gamma}_{LN}) \qquad \text{(A1.4b)}$$

otherwise, birth occurs in a lower level unit (ICU=0). Birth at LNU (LNU==1) would occur if LNU*>ICU* and LNU*>SCU*

$$\tilde{\epsilon}_{LNi} > -\ (\tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi} + X_i' \tilde{\gamma}_{LN}) \qquad \text{(A1.4c)}$$

&

$$\tilde{\epsilon}_{LNi} - \tilde{\epsilon}_{SCi} > -\ \{(\tilde{\alpha}_3 - \tilde{\theta}_3) z_{SCi} + (\tilde{\alpha}_2 - \tilde{\theta}_2) z_{LNi} + (\tilde{\alpha}_1 - \tilde{\theta}_1) z_{ICi} + X_i' (\tilde{\gamma}_{LN} - \tilde{\gamma}_{SC})\}\ ] \qquad \text{(A1.4d)}$$

otherwise birth occurs in a unit of a different level (LNU=0), whilst birth will take place at SCU (SCU=1) if SCU*>ICU* and LNU*<SCU*,

$$\tilde{\epsilon}_{SCi} > -\ (\tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi} + X_i' \tilde{\gamma}_{SC}) \qquad \text{(A1.4e)}$$

&

$$\tilde{\epsilon}_{LNi} - \tilde{\epsilon}_{SCi} < -\{(\tilde{\alpha}_3 - \tilde{\theta}_3) z_{SCi} + (\tilde{\alpha}_2 - \tilde{\theta}_2) z_{LNi} + (\tilde{\alpha}_1 - \tilde{\theta}_1) z_{ICi} + X_i' (\tilde{\gamma}_{LN} - \tilde{\gamma}_{SC})\} \qquad \text{(A1.4f)}$$

otherwise it will occur in a higher level unit.

A1.1 Parametric choice model specification

By assigning a probability distribution to the error terms in 4a and 4b a parametric model of demand for levels of care may be estimated in terms of the right-hand side variables of these equations. In the multinomial logit case, a closed form solution for the choice probabilities, its derivatives and elasticities, exist (Train 2003, p. 61). The 'access price' elasticities of demand for birth at ICU (E_{IC}), LNU (E_{LN}) and SCU (E_{SC}) are as follows:

$$E_{SCk} = (\tilde{\theta}_k (1 - P_{SC}) - \tilde{\alpha}_k P_{LN}) z_k$$

$$E_{LNk} = (\tilde{\alpha}_k (1 - P_{LN}) - \tilde{\theta}_k P_{SC}) z_k$$

$$E_{ICk} = -(\tilde{\alpha}_k P_{LN} + \tilde{\theta}_k P_{SC}) z_k$$

for k= 1,2,3, with P_{SC}, P_{LN} as the multinomial logit choice probabilities of birth at levels 1 and 2, and the access prices (instrumental) variables defined as $z_1 \equiv z_{SC}$, $z_2 \equiv z_{LN}$, $z_3 \equiv z_{IC}$. These are evaluated at sample mean values of the independent variables.

The multinomial logit model has a conveniently simple analytical likelihood function but imposes the assumption that the ratio of choice probabilities between any pair of treatments (e.g. LNU over ICU) is independent from the indirect utility of the third (SCU), i.e. the independence of irrelevant alternatives (IIA) assumption. To relax this assumption, we estimate the parameters of the choice model using the alternative of a multinomial probit distribution that allows for the indirect utility equations 4a and 4b to be correlated. In order to formally identify this model and avoid practical problems of identification (Keane 1992) we impose the exclusion restrictions $\theta_2 = 0, \theta_1 = 0, \alpha_3 = 0$, $\alpha_1 = 0$, $\pi_3 = 0$, $\pi_2 = 0$ on eqs. 3, effectively excluding the travel time to LNU instrument from the indirect utility equation for SCU ($\tilde{\theta}_2 = 0$ in 4a) and viceversa ($\tilde{\alpha}_3 = 0$ in 4b). Notice that the restrictions θ_1= 0 and α_1=0 do not result in any further exclusion restrictions in our differenced demand system of eqs. 4a-4b, or in any change in our a priory expectations on the sign of the coefficients for the included instruments. Thus, in sensitivity analysis of the multinomial probit model, the latent utility system 3 and the indirect utility equations 4a-4c become, respectively:

$$SCU_i^* = \theta_3 z_{SCi} + X_i'\gamma_{SC} + \epsilon_{SCi} \qquad \text{(A1.1.3')}$$

$$LNU_i^* = \alpha_2 z_{LNi} + X_i'\gamma_{LN} + \epsilon_{LNi}$$

$$ICU_i^* = \pi_1 z_{ICi} + X_i'\gamma_{IC} + \epsilon_{ICi}$$

$$\widetilde{SCU}_i^* = \theta_3 z_{SCi} - \pi_1 z_{ICi} + X_i'(\gamma_{SC} - \gamma_{IC}) + \epsilon_{SCi} - \epsilon_{ICi}$$

$$\equiv \theta_3 z_{SCi} + (-\pi_1) z_{ICi} + X_i'\tilde{\gamma}_{SCi} + \tilde{\epsilon}_{SCi} \qquad \text{(A1.1.4a),}$$

$$\widetilde{LNU}_i^* = \alpha_2 z_{LNi} - \pi_1 z_{ICi} + X_i'(\gamma_{LN} - \gamma_{IC}) + \epsilon_{LNi} - \epsilon_{ICi}$$

$$\equiv \alpha_2 z_{LNi} + (-\pi_1) z_{ICi} + X_i'\tilde{\gamma}_{LN} + \tilde{\epsilon}_{LNi} \qquad \text{(A1.1.4b)}$$

$$\widetilde{ICU}_i^* = 0 \qquad \text{(A1.1.4c)}$$

In estimating A1.1.4a-4b we impose the appropriate cross-equation restrictions on the coefficient of z_{ICi} .

Appendix 2 – Endogenous treatment model for each outcome

For all outcomes we estimated endogenous treatment choice models generated by a multinomial logit distribution, we describe this model in section A2.1 below. Alternative endogenous treatment models using the multinomial probit distribution were also estimated for all outcomes except the number of hospital days by level of care (see section A2.2.3 below) and are described in section A2.2.

A2.1 Endogenous multinomial logit treatment choice model

We use the control function approach for estimating Eq. 1' proposed by Debb and Trivedi (Debb and Trivedi 2006),

Following Deb and Trivedi (2006), under the control function approach Eq. 1 becomes

$$Y_i = g\ (\beta_1 SCU_i + \beta_2 LNU_i\ +\ X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi}) +\ v_i \qquad \text{, i.e.}$$

$$g^{-1}(EY_i) =\ \beta_1 SCU_i + \beta_2 LNU_i\ +\ X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi} \qquad \text{(A2.1')}$$

after allowing for a mean zero error, i.e. $Ev_i=0$, with $g^{-1}(EY)$ denoting the link function (logit or probit for mortality, log for costs, and log for days in hospital) evaluated at the mathematical expectation of Y. The terms l_{SCi} and l_{LNi} are unobserved latent utility factors for SCU and LNU (relative to ICU, see eq. 4a & 4b and eq.5), respectively, that affect outcome; the sign of their respective coefficients, e.g. $\lambda_1>0$ and $\lambda_2>0$, would indicate that people whose unobserved characteristics make them to value SCU and LNU more than other people with the same observed characteristics X, tend to have a higher Y. Thus, in this model the two latent factors serve to control for the endogeneity of SCU and LNU in Eq. 1.. Depending on the type of outcome, Eq. A2.1' is estimated by maximising the likelihood of observing the data, given a suitably chosen parametric distribution for v (in our case, binomial for mortality, normal for costs and Poisson for days in hospital).

The model is complemented by the system of differenced latent equations 4a and 4b (4c becomes redundant), which represent the (normalised) indirect utilities of SCU, LNU and NICU, respectively, and the definitions

$$\tilde{\epsilon}_{LNi} \equiv l_{LNi} + e_{LNi} \qquad (5)$$

$$\tilde{\epsilon}_{SCi} \equiv l_{SCi} + e_{SCi}$$

The terms l_{LNi} and l_{SCi} denote the values of unobserved factors affecting the indirect utility of the two options and the infant's outcome Y in Eq A2.1'. The e_{LNi} and e_{SCi} are independently identically distributed error terms that are independent from l_{LNi} and l_{SCi} and whose joint distribution implies a probability of treatment selection given by the function h(.) of linear indices of observed covariates and unobserved factors in 4a & 4b:

$$P(LNU_i,\ SCU_i|z_{SCi}, z_{LNi}, z_{ICi,}X_i, l_{LNi}, l_{SCi}) =\ h(\tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi} + X_i'\tilde{\gamma}_{LN} + l_{LNi}, \tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi} + X_i'\tilde{\gamma}_{SC} + l_{SCi})$$

Thus, the likelihood of observing the sample of data is

$$\int_{i=1}^{N} f(g\ (\beta_1 SCU_i + \beta_2 LNU_i\ +\ X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi}))\ h\big(\tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi} + X_i'\tilde{\gamma}_{LN} + l_{LNi}, \tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi} + X_i'\tilde{\gamma}_{SC} + l_{SCi}\big)\ k(l_{LNi}, l_{SCi}) dl_{SCi}\ dl_{LNi} \qquad \text{(A.2)}$$

where f(.) is the family statistical distribution function and g(.) is the inverse of the link defined in eq. 1' , both of which vary depending on the outcome measure (see A.2.1.1-2.1.3), and k(.) is the

standard normal distribution function for the values of unobserved factors confounding the estimation of causal effects of neonatal care levels on outcomes. Estimation of model parameters may be accomplished by maximising the (simulated) likelihood of the sample of data with multiple values of l_{LNi}, l_{SCi} sampled from k using Halton sequences (Bhat 2001, Debb and Trivedi 2006). The control function approach facilitates estimation of complex non-linear outcome models using simple analytical expressions for the function h(.), which in our case is the multinomial logit. This is accomplished at the expense of imposing the strong assumption that the unobserved confounders l_{LNi}, l_{SCi} enter Eq. 1' linearly.

A2.1.1 Mortality

We estimate the model as a logit link outcome with endogenous multinomial logit treatment (Debb and Trivedi 2006), where eq. A2.1' is the main equation of interest with Y being the unobserved latent propensity to die, Death*:

$$g^{-1}(EDeath_i^*|SCU_i, LNU_i, X_i, l_{SCi}, l_{LNi}) = \beta_1 SCU_i + \beta_2 LNU_i + X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi} \quad \text{(A.2.1.1')}$$

where g^{-1} is the logit link function and f in A.2 is the binomial distribution.

We conducted sensitivity analysis to account for unobserved heterogeneity at the level of the hospital, using an unobserved random term h that varies across hospitals following a normal independent distribution but is fixed within infants born in the same hospital:

$$g^{-1}\left(EDeath_{ij}^*|SCU_{ij}, LNU_{ij}, X_{ij}, l_{SCij}, l_{LNij}\right) = \beta_1 SCU_{ij} + \beta_2 LNU_{ij} + X_{ij}'\delta + \lambda_1 l_{SCij} + \lambda_2 l_{LNij} + h_j \quad \text{(A.2.1.2')}$$

and 4a and 4b now have ij instead of I subscripts and error terms

$$\tilde{\epsilon}_{LNij} \equiv l_{LNij} + \psi_{LN} h_j + e_{LNij}$$

$$\tilde{\epsilon}_{SCij} \equiv l_{SCij} + \psi_{SC} h_j + e_{SCij}$$

with coefficients (loading factors) $\psi_{LN}\ and\ \psi_{SC}$ to be estimated.

A2.1.2 Reimbursement costs and total length of stay

The models for reimbursement costs was estimated using a log normal link function for g^{-1}(.) and a Gaussian family distribution for f(.), where the outcome Y in eq. 1' was the observed reimbursement costs of the infant's inpatient hospital stay. The same specifications applied for the analysis of total length of hospital stay. We also conduct a sensitivity analysis of these models to account for unobserved heterogeneity at the level of the hospital, analogously to the analysis for mortality described in section A2.2.1.

We also estimated a heterogeneous treatment effects model to explore the patterns of self-selection into SCU and LNU treatment according to total costs and costs savings relative to NICU. In brief terms, this consisted in interacting the latent factors l_{SCi} and l_{LNi} with the treatment dummies SCU_i and LNU_i as follows:

$$lnEC_i = (\beta_1 + \lambda_{11} l_{SCi} + \lambda_{12} l_{LNi}) SCU_i + (\beta_2 + \lambda_{21} l_{SCi} + \lambda_{22} l_{LNi}) LNU_i + X_i'\delta + \lambda_1 l_{SCi} + \lambda_2 l_{LNi} \quad \text{(A.2.1.2'')}$$

where ln EC is the natural logarithm of expected costs and, to simplify the expression, we have omitted the fact that the expectation is conditional on the X covariates, the treatment dummies and the latent factors. The equation for length of hospital stay is specified in the same manner. This analysis extends the approach of Card (Card 2001) to multiple treatments using a latent factor set-up as proposed by Aakvik, Heckman and Vytlacil (Aakvik et al. 2005). Further details are available from the authors upon request.

The equations A.2.1.2'', 4a, 4b and 5 are an example of a Potential Outcomes Model, where an individual is observed to have one of three potential outcomes, depending on which treatment option is used:

$Y = Y_3 (1\text{-}SCU\text{-}LNU) + Y_2 LNU + Y_1 SCU$

Where the subscripts indicate the level of care (NICU=3, LNU=2 and SCU=1), and we omit the individual subscripts to simplify notation. In the case of eq. A.2.1.2'' the model takes the form

$Y = Y_3 + (Y_2 - Y_3) LNU + (Y_1 - Y_3) SCU =$

$$\mathrm{E}lnCost_i = X_i'\delta_3 + SCU_i X_i'(\delta_1 - \delta_3) + LNU_i X_i'(\delta_2 - \delta_3)$$

$$+E(\varepsilon_{3i}|ICU{=}1) + LNU_i\{E(\varepsilon_{2i}|LNU{=}1) - E(\varepsilon_{3i}|ICU{=}1)\} + SCU_i\{E(\varepsilon_{1i}|SCU{=}1) - E(\varepsilon_{3i}|ICU{=}1)\} \quad \text{(A2.1.2.2')}$$

with

$$E(\varepsilon_{3i}|ICU{=}1) \equiv \lambda_{1SC} l_{SCi} + \lambda_{2LN} l_{LNi} \quad \text{(A2.1.2.3)}$$

$$LNU_i\{E(\varepsilon_{2i}|LNU{=}1) - E(\varepsilon_{3i}|ICU{=}1)\} \equiv LNU_i(\lambda_{21SC} l_{SCi} + \lambda_{22LN} l_{LNi}) \quad \text{(A2.1.2.4)}$$

$$SCU_i\{E(\varepsilon_{1i}|SCU{=}1) - E(\varepsilon_{3i}|ICU{=}1)\} \equiv SCU_i\,(\lambda_{11SC} l_{SCi} + \lambda_{12SC} l_{LNi}) \quad \text{(A2.1.2.5)}$$

Expression A2.1.2.3 is as in the model by Debb and Trivedi. Expressions A2.1.2.4 and A2.1.2.5 represent the unobserved random coefficients for the effect of the two treatments, which may be correlated with LNU and SCU (endogenous). The six factor loadings λ to be estimated from the data are just identified from the covariation in the two (standard normal) independently distributed heterogeneity latent utility factors in eq. 5, $l = (l_{LN}, l_{SC})$ ~NID(0,1) and the treatment vector $d = (1, SCU, LNU)$ (i.e. 2 X 3=6). (We estimate the correlated coefficients model without the treatment by covariate interactions $SCU_i X_i'(\delta_1 - \delta_3) + LNU_i X_i'(\delta_2 - \delta_3)$ in eq. A.2.1.2.2' due to the high amount of computing time required for a model including these terms, so that differences in potential outcomes are only due to unobserved heterogeneity terms in A2.1.2.2').

In order to validly estimate this model, the instruments must satisfy the condition

$$E(\boldsymbol{\varepsilon}|SCU, LNU, z_{SC}, z_{LN}, z_{IC}) = E(\boldsymbol{\varepsilon}|z_{SCi}, z_{LNi}, z_{ICi}) = \mathbf{0}$$

, which is stronger than condition 2 in the main text. The model is estimated by simulated maximum likelihood (Train 2003). A Wald test for H0: λ_{11}=0, λ_{12}=0, λ_{21} = 0, λ_{22} =0 is a test of the null of exogeneity of treatment effects (i.e. no selection by returns) and λ_{1LN} =0, λ_{2SC} =0 of no selection by severity. This model may be implemented in publicly and commercially available statistical software.

A2.1.3 Number of hospital days spent in each level of care

We estimated the effect of neonatal unit designation on the number of days at British Association of Perinatal and Maternity (BAPM) levels of care 1, 2 and 3 separately, which together accounted for 98% of total LOS (the 'super spell') in our sample. This analysis was implemented using a Poisson endogenous treatment effect model (Debb and Trivedi 2006). The estimated model was as follows:

$$ln\ E(LOS\ level\#_i|X_i)\ = \beta_{1\#}SCU_i + \beta_{2\#}LNU_i + \ X_i{}'\delta_{\#} + \lambda_{1\#}l_{SCi} + \lambda_{2\#}l_{LNi} \quad \text{(A.2.1.4)}$$

where ln E(LOS level#$_i$ |X) is the natural logarithm of the expected number of hospital days at the BAPM level number # by infant i given his or her characteristics X, and other variables are as described before; one model is estimated for each of #=1,2, 3. We present the treatment effects in terms of incidence rate ratios , e.g. in the model for #=1, β_{11} and β_{21} are the proportional increase in the number of days spent at BAPM1 when born at SCU and LNU, respectively, relative to NICU. Presenting results in these relative terms allows us to show the compositional effects of place of birth. To account for endogeneity we estimate this model using the control function approach as implemented by Debb and Trivedi (Debb and Trivedi 2006).

A2.2 Endogenous multinomial probit treatment choice model

A2.2.1 Mortality

We estimated dichotomous regressions of individual infant mortality as a function of place of birth, LNU, SCU or NICU (the reference option) and adjusting covariates,

$$Death_i^* = \beta_{1i}SCU_i + \beta_{2i}LNU_i \ + \ X_i{}'\delta + \varepsilon_i \quad \text{(A.2.2.1)}$$

where Death*$_i$ is a latent continuous variable measuring the death risk of infant i in the follow-up period up to hospital discharge and SCU$_i$ is a binary variable equal to 1 if the neonatal unit of birth of infant i is SCU and 0 otherwise, and LNU is likewise defined for birth in LNU. The term $X_i'\delta$ stands for a linear vector of adjusting covariates (plus a constant) as controls for measured confounding, with their respective coefficients δ. The vector X includes the covariates listed above for birthweight, gestational age, index of multiple deprivation, number of pregnancies, mode of delivery and labour. Since Death* is not observed, but the occurrence, say D=1, or absence, D=0, of death is, the model may be estimated by adopting an observational rule in terms of the value of the error tem ε_i:

$$\begin{aligned} Death_i = \ & 1[Death_i^* = \beta_{1i}SCU_i + \beta_{2i}LNU_i \ + \ X_i'\delta + \varepsilon_i > \tau_\varepsilon] \\ & = 1[\varepsilon_i > \tau_\varepsilon - (\beta_{1i}SCU_i + \beta_{2i}LNU_i \ + \ X_i'\delta)] \end{aligned}$$

which means that an infant's in-hospital death occurs when the latent death propensity is larger than latent death risk threshold τ_ε. The model may thus be estimated by assigning a cumulative distribution function to ε_i , e.g. normal or logistic. Adopting the parametric function implicitly imposing location and scale restrictions on parameters, e.g. in the normal model $\tau_\varepsilon = 0$ and variance of $\varepsilon = 1$.

The probit mortality outcome model with endogenous multinomial probit treatment is estimated by maximising the likelihood of all the permutations of choices among the three possible treatments and the two possible death outcomes (3 x 2 = 6 outcomes) (Roodman 2011). From equations A.2.2.1

and 4a-4b rearranged as Eqs. A1.4a and A1.4b, the likelihoods for mothers whose infants die and survive after giving birth in a NICU hospital are, respectively

$$\Pi_i \Phi_{Ti}\left(-X_i'\delta, -\tilde{\alpha}_3 z_{SCi} - \tilde{\alpha}_2 z_{LNi} - \tilde{\alpha}_1 z_{ICi} - X_i'\tilde{\gamma}_{LN}, -\tilde{\theta}_3 z_{SCi} - \tilde{\theta}_2 z_{LNi} - \tilde{\theta}_1 z_{ICi} - X_i'\tilde{\gamma}_{SC}; \rho_{12}, \rho_{13}, \rho_{23}\right), \text{ for i with Death=1 \& LNU=0 \& SCU=0}$$

$$\Pi_i \Phi_{Ti}\left(X_i'\delta, -\tilde{\alpha}_3 z_{SCi} - \tilde{\alpha}_2 z_{LNi} - \tilde{\alpha}_1 z_{ICi} - X_i'\tilde{\gamma}_{LN}, -\tilde{\theta}_3 z_{SCi} - \tilde{\theta}_2 z_{LNi} - \tilde{\theta}_1 z_{ICi} - X_i'\tilde{\gamma}_{SC}; -\rho_{12}, -\rho_{13}, \rho_{23}\right), \text{ for i with Death=0 \& LNU=0 \& SCU=0}$$

where Φ_T is the tri-variate cumulative normal distribution function with parameters for the correlation between the unobserved error ε_i in the outcome equation (eq. A.2.2.1) and the error term in each of the indirect utility equations 4a ($\tilde{\epsilon}_{SCi}$) and 4b ($\tilde{\epsilon}_{LNi}$) , ρ_{12} and ρ_{13}, respectively; $\rho_{23}$is the correlation between the error terms of these indirect utility equations. Notice that this likelihood is evaluated at SCU=0 and LNU=0 since it only relates to individuals born in NICU. The likelihood functions for mothers whose infants either die or survive after giving birth in LNU and in SCU units are similarly defined after transforming the model latent utility equations into deviations from the equation for those mothers' place of delivery, i.e. LNU or SCU (Train 2003). A Wald test of the null hypothesis Ho: ρ_{12} =0 and Ho: ρ_{13} =0 provides a test of exogeneity of LNU and SCU in equation 1. Furthermore, unlike the multinomial logit model specification of treatment choice, the multinomial probit model allows the relative choice probabilities between any two treatments, say LNU vs. ICU to depend on the characteristics of the third option, say SCU. We allow for a non-zero correlation between random error terms in the indirect utilities of SCU in 4a and LNU in 4b (ρ_{23}).

Sensitivity analysis was conducted using the restricted demand system A1.1.4a and A1.1.4b.

A2.2.2 Reimbursement costs and total length of stay

Since in our dataset all individuals have positive number of inpatient days, the analysis of causal effects of level of care on length of stay and reimbursement cost of neonatal hospital used a log-normal endogenous treatment effects model estimated jointly with a multinomial probit equation for birth in a hospital with level 2 (LNU) or level one (SCU) unit, as a function of a set of exogenous covariates and the travel time instrumental variable. In this case, the estimated model becomes

$$\ln Cost_i = \beta_{1i} SCU_i + \beta_{2i} LNU_i + X_i'\delta + \varepsilon_i \qquad \text{(A2.2.2)}$$

where lnCost is the natural logarithm of reimbursement costs, and the terms on the right hand side are as described for Eq.1 before. The observational rule in terms of the latent equation for SCU and LNU are given by Eqs. A1.4a-A1.4f, and, alternatively, by the equations corresponding to the restricted latent demand system A1.1.4a and A1.1.4b as in the main and sensitivity analyses of mortality. The conditional independence condition required for the validity of the IV now becomes:

$$\boldsymbol{z}_i'\varepsilon_i = \sum_{i=1}^{N} \boldsymbol{z}_i'(lnCost_i - (\beta_{1i} SCU_i + \beta_{2i} LNU_i + X_i'\delta)) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad \text{(A2')}$$

We adopted the assumption that ε and ϵ={ϵ_{sc} ϵ_{ln} ϵ_{ic}} are jointly normally distributed with mean zero, variances σ^2 and {1 1 1}, respectively, and covariance matrix $\boldsymbol{\sigma\rho}$. This model may be estimated by finding the coefficient values that maximise the product of the likelihood functions of the individual cost and unit level of birth observations, following an approach analogous to that described before for the mortality model in section A2.2.1. After transformation, the likelihood function to be maximised is, for infants born in a NICU,

$$\prod_{i=\{ICU=1)} \phi(\ln Cost_i - \left(X_i'\delta - \rho_{12}\sigma \frac{\phi(\tilde{\alpha}\prime z_i + X_i'\gamma_{LN})}{1-\Phi(\alpha\prime z_i + X_i'\gamma_{LN})} - \rho_{13}\sigma \frac{\phi(\tilde{\theta}\prime z_i + X_i'\gamma_{sc})}{1-\Phi(\tilde{\theta}\prime z_i + X_i'\gamma_{sc})}\right))$$

where ϕ is the normal density function and Φ the cumulative normal distribution function, $\tilde{\alpha}'z \equiv \tilde{\alpha}_3 z_{SCi} + \tilde{\alpha}_2 z_{LNi} + \tilde{\alpha}_1 z_{ICi}$ and $\widetilde{\theta}'z \equiv \tilde{\theta}_3 z_{SCi} + \tilde{\theta}_2 z_{LNi} + \tilde{\theta}_1 z_{ICi}$, σ is the variance of ε_i in eq. A2.2.2, and ρ_{12} and ρ_{13} are the correlation between ε_i and $\tilde{\epsilon}_{LNi}$ and $\tilde{\epsilon}_{SCi}$, respectively. As before, the likelihoods of birth at LNU and of birth at SCU are defined in terms of the conditional distribution of the random errors in the propensities of the alternative treatments, differenced relative to LNU* and SCU*, respectively (Train 2009). A Wald test of the null hypothesis Ho: ρ_{12} =0 and Ho: ρ_{13} =0 provides a test of endogeneity of LNU and SCU in equation A2.2.2.

Due to the normal distribution used to analyse log transformed dependent variables in these analyses of LOS and costs, we added the estimated term $\sigma^2/2$ as a correction to the linear indices for obtaining marginal predictions in the original units (Duan 1994). We also estimated censored normal models (results available from authors) to account for censoring from discharge to ward; since results were practically unchanged we only present results of models without censoring adjustment.

The model for the total length of stay simply replaced the dependent variable in Eq. A.2.2.2 with log of the total number of hospital days.

A2.2.3 Number of hospital days spent in each level of care

We did not estimate a multinomial probit specification for this outcome due to the complexity of optimising a likelihood function for an outcome count and a multinomial probit distribution for a multiple endogenous treatment. Thus we only analysed this type of outcomes using the multinomial logit specification described in A2.1.

Appendix 3 - Marginal Treatment Effect analysis

From inspection of the predicted probabilities (propensity scores) of individuals' use of treatments, SCU births had low frequencies at the upper range of propensity scores to NICU and LNU. Since this restricted the feasible range for semiparametric analysis we therefore performed analysis of mortality and neonatal hospital costs by comparing NICU with the LNU and SCU groups combined into a non-NICU treatment group.

In a heterogeneous treatment effects framework an individual is observed to have one of two possible treatments, depending on which treatment option is used:

$$Y = Y_0 (1 - ICU) + Y_1 ICU \quad \text{(A3.1)}$$

where Y_o indicates the potential outcome when born in a non-NICU hospitals and Y_1 the outcome if the infant is born in a NICU, and ICU equals 1 if the infants is born in NICU and equals 0 if born in a non-NICU hospital (we omit the individual subscripts to simply notation). ICU follows the observational rule

$$ICU = 1 \rightarrow ICU_i^* - nonICU_i^* > 0$$

where ICU* is the latent utility equation defined in Eq. 3, and nonICU is the latent utility of birth at a non-NICU correspondingly defined in terms of a linear index of instruments and exogenous variables:

$$nonICU_i^* = \kappa_3 z_{SCi} + \kappa_2 z_{LNi} + \kappa_1 z_{ICi} + X_i' \gamma_{nonIC} + \epsilon_{nonICi}$$

Thus an infant is born in a NICU when

$$\begin{aligned}(\pi_3 - \kappa_3) z_{SCi} + (\pi_2 - \kappa_2) z_{LNi} + (\pi_1 - \kappa_1) z_{ICi} + X_i'(\gamma_{IC} - \gamma_{nonIC}) + (\epsilon_{ICi} - \epsilon_{nonICi}) \\ \equiv \tilde{\pi}_3 z_{SCi} + \tilde{\pi}_2 z_{LNi} + \tilde{\pi}_1 z_{ICi} + X_i' \tilde{\gamma}_{IC} + \tilde{\epsilon}_{ICi} \\ \equiv \tilde{\pi}_3 z_{SCi} + \tilde{\pi}_2 z_{LNi} + \tilde{\pi}_1 z_{ICi} + X_i' \tilde{\gamma}_{IC} - V_i > 0 \quad \text{(A3.2)}\end{aligned}$$

To facilitate interpretation we define the unobserved error $\tilde{\epsilon}_{ICi} \equiv -V_i$, i.e. as the negative of the unobserved 'resistance to use (NICU) treatment' (Carneiro, Heckman and Vytlacil 2010). Moving V_i to the right hand side and evaluating both sides of the inequality by its cumulative distribution function, e.g. the standard normal after normalising the variance of V to 1, the inequality is recast as

$$p(\boldsymbol{z}_i, X_i) \equiv \Phi(\tilde{\pi}_3 z_{SCi} + \tilde{\pi}_2 z_{LNi} + \tilde{\pi}_1 z_{ICi} + X_i' \tilde{\gamma}_{IC}) > \Phi(V_i) \quad \text{(A3.2')}$$

in terms of the propensity score, p(**z**,X), and the percentile rank of the unobserved resistance to give birth at a NICU, Φ(V). The potential outcome framework of Equation A3.5 (Rubin 2005) allows a more general version of a parametric model of costs such as that of Eq. A2.2.2' by allowing two different cost equations, one for each treatment option; i.e. taking the expectation, denoted by E, with respect to the counterpart to A.2.2.2' in terms of Eq. A3.5 is

$$\begin{aligned}&\mathrm{E}lnCost_i = \mathrm{E}[(1 - \mathrm{ICU})\{X_i'\delta_0 + E(\varepsilon_{0i}|\mathrm{ICU}{=}0)\} + ICU\{X_i'\delta_1 + E(\varepsilon_{1i}|ICU{=}1)\}] \\ &ElnCost_i = X_i'\delta_0 + pX_i'(\delta_1 - \delta_0) + E(\varepsilon_{0i}|ICU{=}0) + p\{E(\varepsilon_{1i}|ICU{=}1) - E(\varepsilon_{0i}|ICU{=}0)\} \\ &\qquad \text{(A.1'')}\end{aligned}$$

with p being the probability of NICU birth, i.e. the propensity score on the left-hand side of the inequality in A3.2'. In this model endogeneity may occur in two different forms, one is through the correlation of unobserved factors driving resistance to (NICU) treatment V_i with the underlying baseline severity, $E(\varepsilon_{0i}|ICU{=}0)$; endogeneity may also occur due to unobserved factors in V_i being correlated with unpredictable systematic deviations from average differences in expected costs (negative returns), $\{E(\varepsilon_{1i}|ICU{=}1) - E(\varepsilon_{0i}|ICU{=}0)\}$. In the parametric log normal model with endogenous heterogeneous treatment effects, the estimating equation for Eq. A2.2.2 adopts the form

$$lnCost_i = X_i'\delta_0 + \hat{p_i}X_i'(\delta_1 - \delta_0) + \rho_0\sigma_0 \frac{\phi(\tilde{\pi}\prime z_i + X_i'\tilde{\gamma}_{IC})}{1-\Phi(\tilde{\pi}\prime z_i + X_i'\tilde{\gamma}_{IC})} - \hat{p_i}\{\rho_1\sigma_1 \frac{\phi(\tilde{\pi}\prime z_i + X_i'\tilde{\gamma}_{IC})}{\Phi(\tilde{\pi}\prime z_i + X_i'\tilde{\gamma}_{IC})} + \rho_0\sigma_0 \frac{\phi(\tilde{\pi}\prime z_i + X_i'\tilde{\gamma}_{IC})}{1-\Phi(\tilde{\pi}\prime z_i + X_i'\tilde{\gamma}_{IC})}\}$$

with $\hat{p_i} = \Phi(\tilde{\pi}'z_i + X_i'\tilde{\gamma}_{IC})$. This equation simplifies to

$$lnCost_i = X_i'\delta_0 + \hat{p_i}X_i'(\delta_1 - \delta_0) + (\rho_0\sigma_0 - \rho_1\sigma_1)\phi(\tilde{\pi}'z_i + X_i'\tilde{\gamma}_{IC}) \quad \text{(A3.3)}$$

The pattern of self-selection may thus be estimated by making use of the continuous instrument(s); note that a model that ignores the heterogeneity of treatment effects would fail to distinguish between the two types of self-selection, as it would only estimate a coefficient $(\hat{p}_i(\rho_1\sigma_1 - \rho_0\sigma_0))$, averaged across individuals i. This problem is clearly illustrated by the popular two-step estimator using the inverse Mil's ratio (i.e. $\phi(\tilde{\pi}'z_i + X'\tilde{\gamma}_{IC})/\hat{p}_i$) as an adjusting covariate, which is biased and, although statistical inference on its coefficient may serve to test for self-selection on returns, it has lower power than the heterogeneous treatment effect estimator due to the confounder $\hat{p}_i$.

The marginal treatment effect (MTE) is obtained by taking the derivative of A3.3' with respect to the probability of selection into treatment, i.e. the propensity score $\hat{p}_i$:

$$\begin{aligned}\frac{\partial lnCost}{\partial \hat{p_i}} &= X_i'(\delta_1 - \delta_0) + (\rho_0\sigma_0 - \rho_1\sigma_1)\,\partial\phi\big(\Phi^{-1}(\hat{p_i})\big)/\partial\hat{p_i} \\ &= X_i'(\delta_1 - \delta_0) + (\rho_0\sigma_0 - \rho_1\sigma_1)\;\phi'\big(\Phi^{-1}(\hat{p_i})\big)\,\Phi^{-1\prime}(\hat{p_i}) \\ &= X_i'(\delta_1 - \delta_0) + (\rho_0\sigma_0 - \rho_1\sigma_1)\,\Phi^{-1}(\hat{p_i})\phi\big(\Phi^{-1}(\hat{p_i})\big)\Phi^{-1\prime}(\hat{p_i}) \\ &= X_i'(\delta_1 - \delta_0) + (\rho_0\sigma_0 - \rho_1\sigma_1)\,\Phi^{-1}(\hat{p_i}) \end{aligned} \quad \text{(A3.4)}$$

where the first step makes use of the fact that the liner index $\tilde{\pi}'z_i + X_i'\tilde{\gamma}_{IC} = \Phi^{-1}(\hat{p_i})$, the inverse of the cumulative normal distribution function, the second step is obtained by the chain rule, the third step results from the exponential form of the density normal function, and the final equation from the chain rule and definition of the inverse function $\Phi^{-1\prime}(\hat{p_i}) = 1/\phi\big(\Phi^{-1}(\hat{p_i})\big)$.

Coefficient estimates may be interpreted in analogous form to those of a linear probability model of mortality, as the dependent variable, instead of costs. There will be unobservable positive, negative and no NICU self-selection on severity of condition (or costs) if $\rho_0 > 0$, $\rho < 0$, $\rho = 0$, respectively; and positive, negative or no self-selection on returns in terms of mortality risk reduction (or increased healthcare costs) when $\rho_1 - \rho_0$ is positive, negative or zero, respectively. For example, a positive self-selection on severity means in this context that unobserved characteristics that make infants more likely to be born at a NICU hospital predispose them also to worse health outcomes (i.e. higher costs) than the average infant. A selection on returns means that unobserved characteristics that make infants more likely to be born in a NICU also characterise those NICU births with higher

expected additional costs to the NHS, relative to birth in a lower designation hospital. Notice that the second term on the right hand side of Eq A3.4 is the product of a coefficient and the unobserved individual-specific resistance to deliver at NICU at the margin (i.e. where V=p), $V_i=\Phi^{-1}(\hat{p}_i)$ so that the term - $(\sigma_1 \rho_1 - \sigma_0 \rho_0)V_i$ represents additional costs among individuals with unobserved resistance to birth at NICU of V_i (Carneiro, Heckman and Vitlacyl 2010).

We also estimate MTE in a linear probability model of mortality where the observed binary Death indicator replaced log costs as the dependent variable in Eq. A3.3 . Furthermore, we relax the parametric assumption for the estimation of MTE on the linear models for a) reimbursement costs and b) probability of mortality in the binary NICU vs non-NICU cases, by estimating a semi-parametric model using the local IV estimator (Heckman and Vytlacil 1999), a two-step approach to local polynomial regression of degree 4 which allows for non-monotonic patterns of selection with the unobserved resistance to treatment V defined above (Brave and Walstrum 2014). This model still requires the assumption that the resistance to treatment term V enters linearly in the treatment equation so that its region of common support may be applied across all the possible subgroups defined by unique combinations of covariate values X.

Briefly, the formula for the semiparametric estimator of the MTE is, in the example for reimbursement costs (the analysis of mortality is analogous to this example, using a linear probability of death specification)

$$\frac{\partial lnCost}{\partial \hat{p}} = X'(\delta_1 - \delta_0) + \partial K(\hat{p})/\partial \hat{p} \qquad \text{(A3.5)}$$

where, the last term is a general function that is estimated using nonparametric techniques for local derivatives. We use the estimation approach known as local instrumental variables (LIV), which consists on first running local linear regressions of X, Xp, and lnCost on p at every observed value of $\hat{p}$ to obtain estimated residuals $\hat{e}_X$, $\hat{e}_{Xp}$ and $\hat{e}_{lnCost}$; $\hat{e}_{lnCost}$is then regressed on $\hat{e}_X$ and $\hat{e}_{Xp}$ to obtain an estimate of δ_0, $(\delta_1 - \delta_0)$; the rest of the parameters of the MTE are estimated from a local polynomial regression of

$$\widetilde{lnCost} \equiv lnCost - X'\hat{\delta}_0 - X'(\hat{\delta}_1 - \hat{\delta}_0)p$$

on the common support of p to obtain $\partial K(\widehat{p_i})/\partial \widehat{p_i}$. In a separate analysis we also estimate a flexible parametric model by approximating the error correction term using a polynomial of degree 4. Standard errors and 95% confidence intervals for the MTE are obtained using bootstrapping with 50 resamples, with this procedure repeated at each resample. The MTEs are estimated conditional on the mean values of X in the sample and the standard errors do not reflect sampling variation in X (see Brave and Walstrum 2014 for further details).

Appendix 4 – Stata code to implement the main analyses

The following code implements the main analyses in Stata 14 using cmp command (Roodman 2014) for the multinomial probit specification, the official Stata command gsem for the multinomial logit model and the margte (Brave ad Walstrum 2014).

A4.1 In-hospital mortality

A4.1.1 Endogenous multinomial probit treatment (with exclusion restrictions)

```
local covars = "gestation_weeks Sqgestation_weeks KgBirthweight  SQKGBirth fetus_number Female quintileIMD* MDEmergCesarianNotLabour MDEmergCesarianLabour Vaginalnonspont ElectiveSection Unknown"

constraint 1 [_outcome_2_3]mindisttimescu=0

constraint 2 [_outcome_2_4]mindistimehdu=0

constraint 3 [_outcome_2_3]mindisttimeicu=[_outcome_2_4]mindisttimeicu=0

cmp (died = i.SCU i.LNU `covars') (mtreat = mindisttimeicu mindistimehdu mindisttimescu `covars') if lsoa_has_all_unit==1, ind($cmp_probit $cmp_mprobit) quietly vce(cluster place_of_birth) constr(1 2 3)

test [_outcome_2_3 ]:mindisttimeicu mindisttimescu mindistimehdu

test [_outcome_2_4 ]:mindisttimeicu mindisttimescu mindistimehdu

margins, dydx(SCU) predict(pr equation(died)) force at(LNU==0)

margins, dydx(LNU) predict(pr equation(died)) force at(SCU==0)
```

A4.1.2 Endogenous multinomial logit treatment

```
gsem (died <- 1.LNU 1.SCU `covars' L2 L3, family(binomial) link(logit)) (2.mtreat <- mindisttimeicu mindistimehdu mindisttimescu `covars' L2@1, mlogit) (3.mtreat <- mindisttimeicu mindistimehdu mindisttimescu `covars' L3@1, mlogit) if lsoa_has_all_unit_types_closest==1, var(L2@1 L3@1) nocaps latent(L2 L3) vce(cluster place_of_birth) cov(L2*L3@0) startvalues(zero)

margins, dydx(SCU) expression(predict(equation(died))) force at(LNU==0)

margins, dydx(LNU) expression(predict(equation(died))) force at(SCU==0)

margins SCU, expression(predict(equation(died))) force at(LNU==0)

margins LNU, expression(predict(equation(died))) force at(SCU==0)

test [2.mtreat]:mindisttimeicu mindisttimescu mindistimehdu

test [3.mtreat]:mindisttimeicu mindisttimescu mindistimehdu
```

A4.2 Reimbursement costs and length of hospital stay

Code is provided below for reimbursement costs only, length of hospital stay is implemented in the same way for the respective outcome variable.

A4.2.1 Endogenous multinomial probit treatment (with exclusion restrictions)

```
constraint 1 [_outcome_2_3]mindisttimescu=0

constraint 2 [_outcome_2_4]mindistimehdu=0

constraint 3 [_outcome_2_3]mindisttimeicu=[_outcome_2_4]mindisttimeicu=0


cmp (lTotalHRG = i.SCU i.LNU `covars') (mtreat = mindisttimeicu mindistimehdu mindisttimescu `covars') if lsoa_has_all_unit==1, ind($cmp_cont $cmp_mprobit) quietly vce(cluster place_of_birth) constr(1 2 3)

test [_outcome_2_3 ]:mindisttimeicu mindisttimescu mindistimehdu

test [_outcome_2_4 ]:mindisttimeicu mindisttimescu mindistimehdu

margins SCU, expression(exp(predict(equation(lTotalHRG))+(exp([lnsig_1]_b[_cons])^2)/2)) force at(LNU=0)

margins LNU, expression(exp(predict(equation(lTotalHRG))+(exp([lnsig_1]_b[_cons])^2)/2)) force at(SCU=0)

margins, dydx(SCU) expression(exp(predict(equation(lTotalHRG))+(exp([lnsig_1]_b[_cons])^2)/2)) force at(LNU==0)

margins, dydx(LNU) expression(exp(predict(equation(lTotalHRG))+(exp([lnsig_1]_b[_cons])^2)/2)) force at(SCU==0)
```

A4.2.2 Endogenous multinomial logit treatment

```
gsem (lTotalHRG <- 1.LNU 1.SCU `covars' L2 L3, family(gaussian) link(identity)) (2.mtreat <- mindisttimeicu mindistimehdu mindisttimescu `covars' L2@1, mlogit) (3.mtreat <- mindisttimeicu mindistimehdu mindisttimescu `covars' L3@1, mlogit) if lsoa_has_all_unit_types_closest==1, var(L2@1 L3@1) nocaps latent(L2 L3) vce(cluster place_of_birth) cov(L2*L3@0)

margins SCU, expression(exp(predict(equation(lTotalHRG))+_b[var(e.lTotalHRG):_cons]/2)) force at(LNU==0)

margins LNU, expression(exp(predict(equation(lTotalHRG))+_b[var(e.lTotalHRG):_cons]/2)) force at(SCU==0)

margins, dydx(SCU) expression(exp(predict(equation(lTotalHRG))+_b[var(e.lTotalHRG):_cons]/2)) force at(LNU==0)

margins, dydx(LNU) expression(exp(predict(equation(lTotalHRG))+_b[var(e.lTotalHRG):_cons]/2)) force at(SCU==0)

test [2.mtreat]:mindisttimeicu mindisttimescu mindistimehdu

test [3.mtreat]:mindisttimeicu mindisttimescu mindistimehdu
```

A4.3 Marginal treatment effects for binary NICU vs. non-NICU treatment comparison

A4.3.1 In-hospital mortality

margte died `covars' if lsoa_has_all_unit==1, treatment(ICU mindisttimeicu mindisttimescu mindistimehdu `covars') semiparametric polynomial(4)

A4.3.2 Reimbursement costs

margte lTotalHRG `covars' if lsoa_has_all_unit==1, treatment(ICU mindisttimeicu mindisttimescu mindistimehdu `covars') semiparametric polynomial(4)

Appendix 5. Distribution of sample characteristics by travel time to SCU and LNU

Table A5.1 Sample characteristics by travel time to LNUs and SCUs (% unless stated otherwise)

	Travel time to LNU (N=11,037)			Travel time to SCU (N=11,037)		
	Lower tertile N=3,377	Medium tertile N=3,696	High tertile N=3,964	Lower tertile N=3,672	Medium tertile N=3,475	High tertile N=3,890
Died	8.59	8.63	8.37	8.36	8.66	8.56
Discharged home	86.40	86.52	86.80	86.79	86.35	86.60
Discharged ward	1.69	1.49	1.84	1.63	1.81	1.59
Last record: transferred to another hospital/unit	3.05	3.30	2.75	2.97	3.05	3.05
Unknown destination	0.27	0.05	0.23	0.24	0.11	0.18
Gestational age at birth (weeks), mean (SD)	28.44 (2.35)	28.48 (2.33)	28.45 (2.32)	28.38 (2.38)	28.44 (2.32)	28.53 (2.30)
Birthweight (kg), mean (SD)	1.19 (0.38)	1.20 (0.39)	1.21 (0.39)	1.19 (0.38)	1.20 (0.38)	1.22 (0.39)
Foetus 2+	25.64	27.35	27.88	28.51	26.47	26.09
Female sex	45.54	45.21	46.47	44.88	46.59	45.86
Residence: Most deprived quintile[1]	37.13	34.55	31.89	36.52	33.52	33.14
Residence: 2nd most deprived quintile[1]	23.08	22.48	20.41	23.53	24.11	20.18
Residence: 3-5 least deprived quintile[3]	37.78	42.96	47.70	39.95	42.36	46.68
Caesarean delivery	52.80	50.05	48.59	50.27	51.71	49.25
Spontaneous vaginal	36.19	36.15	38.02	35.21	37.35	37.89
Unknown delivery mode	3.26	4.22	0.00	5.42	2.73	0.00
Delivery at NICU	28.87	54.11	66.09	51.42	46.50	53.75
Delivery at LNU	68.61	35.06	17.20	22.19	49.01	45.68
Delivery at SCU	2.52	10.82	16.70	26.39	4.49	0.56
Delivery at high volume[2]	20.22	29.22	15.49	23.53	19.51	21.46

[1] Ranked by the index of multiple deprivation of residential postcode. [2] Defined as born in hospital delivering more than 100 infants with <1500 g birthweight per year during the study period. SD: Standard deviation.

Table A5.2 Sample characteristics by travel time to closest NICU, LNU, SCU in infants born in a London hospital (% unless stated otherwise)

	Travel time to ICU (N=1,829)			Travel time to LNU (N=1,829)			Travel time to SCU (N=1,829)		
	Lower tertile N=653	Medium tertile N=600	High tertile N=576	Lower tertile N=539	Medium tertile N=623	High tertile N=667	Lower tertile N=689	Medium tertile N=524	High tertile N=616
Died	7.96	6.00	7.29	5.57	6.42	8.99	7.26	7.63	6.49
Discharged home	86.68	91.17	87.98	90.15	89.25	86.64	87.94	88.17	89.59
Discharged ward	2.60	1.17	1.57	0.93	1.60	2.70	2.33	1.53	1.46
Last record: transferred to another hospital/unit	2.30	1.33	2.96	2.79	2.25	1.65	2.33	2.29	1.95
Unknown destination	0.46	0.33	0.17	0.56	0.48	0	0.14	0.38	0.49
Gestational age at birth (weeks), mean (SD)	28.31 (2.40)	28.60 (2.29)	28.14 (2.34)	28.39 (2.36)	28.31 (2.37)	28.36 (2.33)	28.31 (2.32)	28.42 (2.34)	28.34 (2.41)
Birthweight (kg), mean (SD)	1.17 (0.38)	1.19 (0.37)	1.13 (0.37)	1.19 (0.37)	1.14 (0.37)	1.17 (0.39)	1.17 (0.38)	1.16 (0.36)	1.17 (0.39)
Foetus 2+	29.55	31.50	26.91	28.94	28.41	30.58	30.48	28.82	28.57
Female sex	44.26	44.50	47.05	47.68	43.82	44.53	43.83	47.52	44.80
Residence: Most deprived quintile[1]	44.26	23.00	22.22	29.87	31.62	29.53	33.67	29.20	27.60
Residence: 2nd most deprived quintile[1]	35.68	35.00	27.43	35.81	34.51	28.93	27.58	34.35	37.5
Residence: 3-5 least deprived quintile[3]	20.06	42.00	50.35	34.32	33.87	41.53	38.75	36.45	34.90
Caesarean delivery	52.83	55.17	53.65	54.73	54.73	52.32	51.09	54.96	56.01
Spontaneous vaginal	35.22	35.17	35.24	35.62	35.15	34.93	36.43	33.78	35.06
Unknown delivery mode	1.99	1.17	0.00	1.67	2.09	0.00	1.45	1.91	0
Delivery at NICU	70.60	31.67	36.28	38.77	44.30	56.22	52.54	49.05	39.12
Delivery at LNU	24.96	49.67	46.87	57.70	46.71	19.34	22.21	40.08	59.74
Delivery at SCU	4.44	18.67	16.84	3.52	8.99	24.44	25.25	10.88	1.14
Delivery at high volume[2]	58.50	26.17	30.73	33.77	37.24	45.28	45.28	34.35	36.36

[1] Ranked by the index of multiple deprivation of residential postcode. [2] Defined as born in hospital delivering more than 100 infants with <1500 g birthweight per year during the study period. SD: Standard deviation.

Appendix 6 – Main results for extremely preterm births

Table A6.1 Causal effect of birth in lower level hospitals (LNU & SCU relative to ICU) on mortality in infants born less at <28 weeks gestation

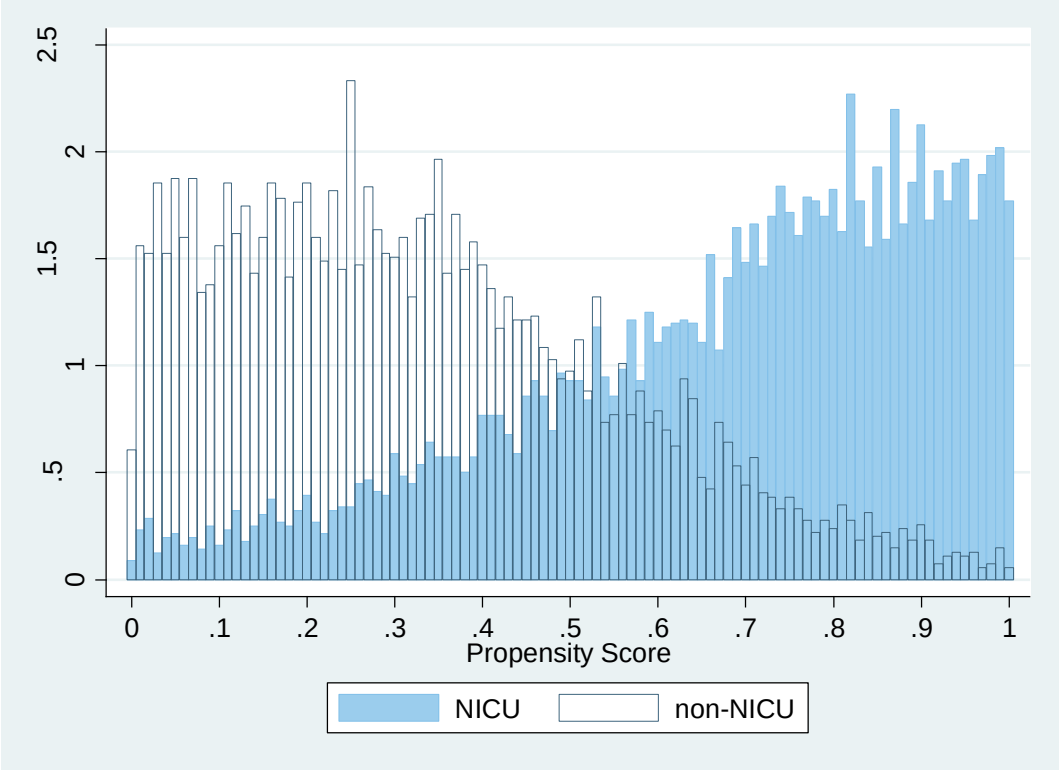
	Naïve	**IV**
	logit regression	**Logit with endogenous multinomial logit treatment**
Birth at LNU	0.027 (0.021)	0.030 (0.041)
Birth at SCU	0.064* (0.033)	0.011 (0.089)
Instrument strength: Wald F test statistic (3 degrees of freedom)	N/A	LNU equation: 79*** SCU equation: 90***
N	3,394	3,394
Hausman test z statistic of H0: no endogeneity LNU treatment variable	N/A	-0.13
Hausman test z statistic of H0: no endogeneity SCU treatment variable	N/A	0.56
Test z statistic Ho: valid over-identifying restriction of minimum travel time to NICU	N/A	0.28

*Controlled covariates: Age and age squared at birth, birthweight, birthweight squared, sex, deprivation of residence, mode of delivery, foetus no. N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01.N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01. Statistical inferences based on robust standard errors adjusting for clustering of observations by hospital.*

Appendix 7 – Results on Marginal Treatment Effects (MTEs)

Allowing for heterogeneous treatment effects and exploiting the continuous instruments, we analyse the causal effect of birth at level 3 relative to a lower-designation hospital, by level of unobserved resistance to treatment (which for mothers in the margin, i.e. indifferent between delivering in NICU vs. non-NICU, is equal to the propensity score). This analysis adopts a linear probability model of mortality with a binary endogenous treatment instrumented by the same three instruments as before. Naturally, the area of common support of the resistance to treatment V, presented in Figure A7.1, shows that NICU-born infants are concentrated on the upper half of the propensity score (i.e. the probability of birth in a hospital with NICU) ranking, whereas non-NICU born infants are concentrated on the lower half. There is overlap in the distribution of intervention (NICU births) and control (non-NICU) subjects by level of the propensity score across the whole range of the probability ranking, which allows us to estimate marginal treatment effects (MTEs) semi-parametrically for individuals ranked in their unobserved resistance to NICU from a propensity score close to 1 (most resistant or least inclined) to near 0 (least resistance or most inclined to delivering in NICU).

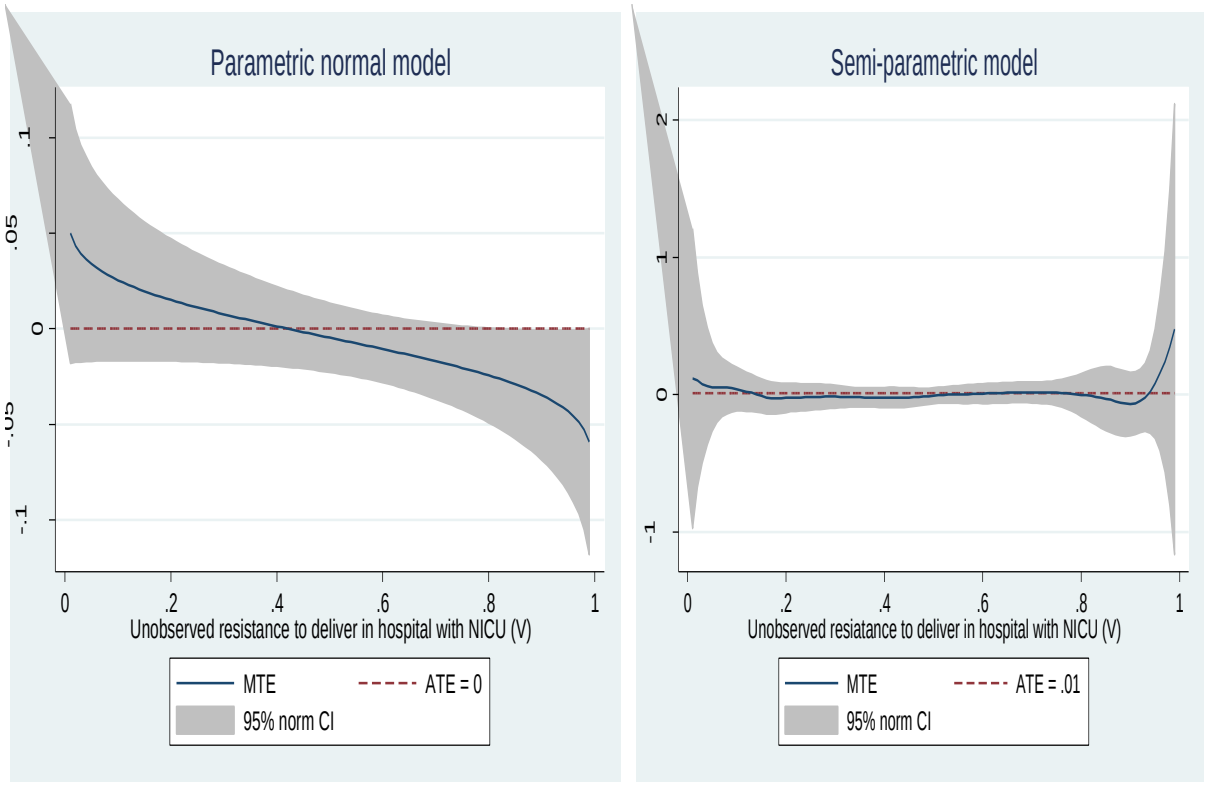
Figure A7.1. Frequency of propensity score by treatment status



The parametric normal model produced average treatment effects (ATE) estimates for in-hospital mortality at the mean of the exogenous covariates of -0.005 (95% -0.020, 0.010). In terms of selection patterns, the estimates were $\sigma_0\rho_0 = 0.015$ (p=0.097), and $\sigma_1\rho_1 - \sigma_0\rho_0 = -.023$ (p=0.054). The ATE and MTEs for individuals with different resistances to be born at NICU are presented in Figure A7.2, under the normal and the alternative semi-parametric model specifications. Under the parametric normal model, while the average treatment effect is zero across all infants, those within

the top 20% unobserved resistance to birth at a NICU have positive expected mortality risk reductions. However, the MTEs under the flexible semi-parametric specification do not appear to vary with the level of unobserved composite factors that discourage birth at a NICU hospital.

Figure A7.2. Marginal Treatment Effects on mortality of level 3 vs. lower designation hospital



Under the parametric normal MTE model, NICU reduced neonatal hospital reimbursement costs relative to hospitals of lower designation by 2.9% (Average treatment effect; 95% CI: 0.088, 0.031). Further, non-observably severe or complex (i.e. more costly) patients appear to positively select into NICU, $\sigma_0\rho_1 = 0.071$ (p=0.001) but there is no apparent selection by cost savings $\sigma_1\rho_1 - \sigma_0\rho_0 = -0.054$ (p=0.17). Semi-parametric analysis reveals, however, that some of the individuals with the highest propensity to deliver at NICU (i.e. in the 20 to 40 percentile range of the ranking) have the highest excess costs relative to a non-NICU birthplace; see Figure 1. For other individuals there is no evidence of NICU resulting in costs different to those of non-NICU care.

Appendix 8 – Results on length of hospital stay and reimbursement costs for births at <32 weeks

Based on the IV multinomial logit model, the estimated total duration of the infant's hospital stay including hospital transfers (i.e. the 'super spell') of an infant born in NICU or LNU was 66 days, and 67 days if birth were to take place in a SCU. These differences were not statistically significant, however (SCU vs NICU, p=0.72; LNU vs. NICU, p=0.76; Table A8.1). The reimbursement cost of birth in a NICU, was £42,776 per infant, and the difference from birth at a lower level unit was imprecisely estimated (NICU minus LNU, -£2078 [95% CI: -5551, 1396]; NICU minus SCU, -£444 [-4690, 3802]; Table A8.1). Hausman test statistics are consistent with the presence of treatment endogeneity.

Despite no evidence of cost differences between neonatal unit designations overall, there is evidence of heterogeneous treatment effects of LNU relative to NICU according to the correlated coefficients model (Table A8.1). There is negative self-selection by severity in the sense that the more individuals value birth at LNU (relative to birth at NICU) the lower their level of costs (i.e. in column 8, $\lambda_{2LNU} < 0$, $p<0.01$). Further, there is positive self-selection by losses, since the stronger the preference for LNU the higher the incremental cost of birth at LNU relative to birth at a NICU ($\lambda_{22LNU} > 0$).There is no apparent selection by severity or selection by returns for SCU ($p>0.10$ for all the respective coefficients). Notice that since the latent utility of birth at LNU is defined relative to the utility at birth in ICU (eq. 4a, 4b), the previous statements about the value of birth at LNU may be equivalently phrased as positive self-selection by severity into NICU and positive self-selection by losses into NICU. Results for LOS mirror those described for costs (and may be compared as the estimated lambdas represent the covariance of the latent utility errors with the log costs and log LOS and are therefore scale-free), with the exception that patients with stronger preference for SCU have shorter inpatient hospital super spells (selection by severity; $\lambda_{1SCU} = -0.28$, $p<0.05$, Table A8.1, column 4). Sensitivity analyses show that the results for reimbursement cost and LOS results in Table A8.1 are robust to inclusion of mode of delivery covariates (available from the authors).

In the above analyses it must be kept in mind that the LOS and cost differences between neonatal care designations is likely to reflect the effects of differences in mortality between these treatment options. In particular if the small increase in mortality observed with SCU designation relative to NICU in the naïve analysis were to be true, the differences in LOS between the two groups would underestimate the expected difference in LOS for infants with the same health outcomes (see Table 3). A similar consideration would apply to the analysis of costs, although in this case the effect might have been partly offset by an increase in per diem costs due to more use of intensive care (BAPM 1) in SCU due to the excess deaths relative to NICU (Figure 2). Therefore, these analyses of costs and LOS, which happened to be consistent with no average treatment effect, ought to be interpreted in the light of the corresponding estimated effects on mortality.

Table A8.1 Marginal effects on LOS and reimbursement costs of birth in LNU & SCU relative to ICU

	LOS (log normal)				Reimbursement costs (log normal)			
	Naïve	IV endogenous multinomial treatment		Correlated random coefficients	Naïve	IV endogenous multinomial treatment		Correlated random coefficients[1]
	OLS	Multi-nomial probit	Multi-nomial logit	Multi-nomial logit	OLS	Multi-nomial probit	Multi-nomial logit	Multi-nomial logit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth at LNU	-2.03 (1.97)	-0.50 (2.62)	0.64 (2.69)	1.79 (2.54)	-709 (1249)	1293 (1773)	2078 (1772)	2044 (1614)
Birth at SCU	-3.67 (2.30)	3.22 (3.45)	1.00 (3.24)	4.02 (3.05)	-2073 (1520)	2102 (2303)	444 (2166)	2346 (2041)
λ_{2LNU}			-0.07 (0.06)	-0.27** (0.11)			-0.11** (0.06)	-0.31*** (0.12)
λ_{22LNU}				0.32** (0.12)				0.34** (0.13)
λ_{12LNU}				0.02 (0.15)				0.05 (0.17)
λ_{1SCU}			-0.11 (0.07)	-0.28** (0.11)			-0.09 (0.08)	-0.16 (0.26)
λ_{21SCU}				0.20 (0.12)				0.09 (0.23)
λ_{11SCU}				0.11 (0.12)				0.01 (0.25)
ρ_{12}		-0.03				-0.06		
ρ_{13}		-0.08**				-0.08**		
ρ_{23}		0.85**				0.85**		
Instrument strength: Wald F test statistic (3 degrees of freedom)	N/A	LNU equation: 199*** SCU equation: 40***	LNU equation: 171*** SCU equation: 150***	LNU equation: 171*** SCU equation: 150***	N/A	LNU equation: 198*** SCU equation: 40***	LNU equation: 171*** SCU equation: 150***	LNU equation: 170*** SCU equation: 149***
N	11,037	11,037	11,037	11,037	11,037	11,037	11,037	11,037
Hausman test z statistic of H0: no endogeneity Birth at LNU	N/A	-0.92	-1.09	-2.55**[2]	N/A	-1.61	-1.97**	-2.62***[2]
Hausman test z statistic of H0: no endogeneity Birth at SCU	N/A	-2.34**	-1.53	-2.60**[2]	N/A	-2.16**	-1.16	0.62 [2]
z statistic of H0: no correlation	N/A	2.25**	Implicit assumption H0 is true	Implicit assumption H0 is true	N/A	2.20**	Implicit assumpti	Implicit assumpti

between utility equations (IIA)							on H0 is true	on H0 is true
Test z statistic Ho: valid over-identifying restriction of minimum travel time to NICU	N/A	-0.69	-0.99	N/A	N/A	-1.29	1.68	N/A

*Models were estimated with logarithm-transformed dependent variables; marginal effects were calculated by exponential back transformation adjusted for conditional variance of log dependent (Duan 2012). Controlled covariates: Age and age squared at birth, birthweight, birthweight squared, sex, deprivation of residence, foetus no. N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01.N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01. Statistical inferences based on robust standard errors adjusting for clustering of observations by hospital. 1. The model omitted the treatment indicators by covariate interactions due to time expense required to estimate the fully heterogeneous treatment model. 2. In order to keep comparability with other statistics in the same row, this statistic is not for testing endogeneity of treatment, but for whether there is selection by severity: under treatment heterogeneity as in the correlated coefficients model endogeneity is the combined results of selection on severity and selection on returns (see eq. A2.1.2.2', and eq. A2.1.2.3-5 in section A2.1.2).*

Appendix 9 – Results on inpatient days by level of care for VPT births

Table A9.1 Marginal effects on inpatient days by level of care in LNU & SCU relative to ICU

	Naïve Negative binomial regression				**IV Negative binomial regression with endogenous multinomial treatment[1]**			
	Total length of stay	**BAPM Level 1**	**BAPM Level 2**	**BAPM Level 3**	**Total length of stay**	**BAPM Level 1**	**BAPM Level 2**	**BAPM Level 3**
Birth at LNU	-0.36 (1.10)	0.48 (0.80)	0.68 (0.93)	-1.30 (0.84)	-0.15 (1.09)	4.88*** (0.84)	0.81 (2.12)	-1.35 (0.82)
Birth at SCU	-0.11 (1.74)	0.93 (1.00)	-0.12 (1.74)	-0.34 (1.11)	-0.07 (1.74)	0.75 (2.23)	1.67 (1.19)	0.32 (1.12)
λ_{1LNU}					-0.01** (0.00)	-0.45*** (0.05)	-0.09 (0.07)	-0.00 (0.00)
λ_{2SCU}					-0.00 (0.00)	0.06 (0.24)	-0.10 (0.07)	-0.00 (0.00)
Instrument strength: Wald F test statistic (3 degrees of freedom)					LNU eq.: 172*** SCU eq.: 150***	LNU eq.: 169*** SCU eq.: 155***	LNU eq.: 172*** SCU eq.: 151***	LNU eq.: 171*** SCU eq.: 150***
N	11,037	11,037	11,037	11,037	11,037	11,037	11,037	11,037
Hausman test z statistic of H0: no endogeneity LNU						-8.35***	-1.26	0.81
Hausman test z statistic of H0: no endogeneity SCU						0.26	-1.42	-0.90
z statistic: H0: lnψ=0 (variance of over-dispersion term)	-26.28***	-3.53***	4.23***	-6.92***		-9.65***	3.85***	6.92***

*Models were estimated with logarithm link function and mean dispersion. Controlled covariates: Age and age squared at birth, birthweight, birthweight squared, sex, deprivation of residence, foetus no. N/A: Not applicable. *p< 0.10 ** p<0.05 ***p<0.01. Statistical inferences based on robust standard errors adjusting for clustering of observations by hospital. For details see section A2.1.3.*

Despite no apparent differences in the length of inpatient stay across treatments, there was an increase of 5 days of intensive care (level 1) with LNU relative to NICU. No other differences were detected between treatments across levels of care.

The negative binomial regression specification used in the analysis extends the Poisson model by an over-dispersion term. Statistical tests suggest that the data were inconsistent with the Poisson

model ($\ln\psi=0$ was rejected at conventional levels of statistical significance) under both the naïve and multinomial endogenous models of inpatient hospital days for all levels of care.