LSE Research Online

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Eric B. Schneider

Fetal health stagnation: have health conditions in utero improved in the United States and western and northern Europe over the past 150 years?

Article (Accepted version)
Refereed

Original citation:
Schneider, Eric B. (2017) *Fetal health stagnation: have health conditions in utero improved in the United States and western and northern Europe over the past 150 years?* Social Science and Medicine . ISSN 0277-9536

DOI: 10.1016/j.socscimed.2017.02.018

Reuse of this item is permitted through licensing under the Creative Commons:

© 2017 Elsevier
CC BY-NC-ND 4.0

This version available at: http://eprints.lse.ac.uk/69510/

Available in LSE Research Online: February 2017

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may freely distribute the URL (http://eprints.lse.ac.uk) of the LSE Research Online website.

http://eprints.lse.ac.uk

FETAL HEALTH STAGNATION: HAVE HEALTH CONDITIONS IN UTERO IMPROVED IN THE US AND WESTERN AND NORTHERN EUROPE OVER THE PAST 150 YEARS?

Dr Eric B. Schneider[1]

Assistant Professor
Economic History Department, LSE

Houghton Street
London, WC2A 2AE
United Kingdom

e.b.schneider@lse.ac.uk

7 February 2017

1 The author wishes to thank Vellore Arthi, Samantha Rawlings and Torvid Kiserud for helpful comments on the paper along with attendees at the 2014 Beyond Birth Weight Workshop at the University of Sussex, the demography seminar at Umeå University, the 2015 NBER Cohort Studies Meeting in Boston, the Economic History Society Meeting in Wolverhampton, the Third CEPR Economic History Symposium in Oslo, the European Historical Economics Society Conference, The Ninth World Congress of the International Society for the Developmental Origins of Health and Disease, and the Health and Skills Formation Workshop at the University of Essex. The usual disclaimer applies. Funding from ESRC grant ES/L010267/1 is gratefully acknowledged.

Abstract

Many empirical studies have shown that health conditions *in utero* can have long lasting consequences for health across the life course. However, despite this evidence, there is no clear consensus about how fetal health has changed in the very long run. This paper analyses historical birth weights and perinatal mortality rates to construct a coherent picture of how health conditions *in utero* have changed over the past 150 years. In short, the evidence suggests that fetal health has been relatively stagnant. Limited evidence on birth weights shows that they had already reached their current levels in North America and Northern and Western Europe by the late nineteenth century, and they have changed very little in between. Perinatal mortality rates have fallen dramatically since the late 1930s, but this decline was mainly caused by improvements in intrapartum treatments after the introduction of Sulfa drugs and antibiotics. Thus, the health benefits associated with the perinatal mortality decline were concentrated among those at risk and did not influence the population at large. Finding stagnant fetal health during a period when many other indicators of health improved dramatically is provocative and suggests two conclusions: either fetal health did not improve or the indicators used to measure fetal health, indicators still widely used today, may not accurately capture all aspects of health *in utero*. If fetal health has been stagnant, then better conditions *in utero* cannot explain cohort improvements in life expectancy over the twentieth century. If the indicators of fetal health are problematic, then researchers must move beyond birth weight and perinatal mortality to understand how developmental plasticity based on the prenatal environment influences later life health.

Keywords

Health history; fetal health; birth weight; perinatal mortality

Introduction

Many empirical studies have shown that health conditions experienced by foetuses *in utero* have significant long-lasting health consequences. Babies exposed to poor conditions are at higher risk for heart disease, stroke and diabetes in later life and have lower lifetime earnings and educational attainment and greater disability than healthier cohorts (Almond and Currie, 2011; Barker, 1997; Conley *et al.*, 2003; Figlio *et al.*, 2014; Godfrey *et al.*, 2007). Studies to date have established a causal link between fetal health and later life health, but they have often relied on exogenous shocks to cleanly identify causal links. Thus, there have been relatively few studies that attempt to explain how fetal health has changed over time.

The period between 1860 and the present has been a period of epidemiological transition where many standard indicators of human health have improved dramatically around the world. Crude death rates, child mortality, infant mortality and stillbirth rates have fallen. Life expectancy and average adult height have increased. Western Europe and North America led these trends with the rest of the world following suit in the second half of the twentieth century. The earliest aspects of the mortality decline occurred apart from modern medical science before the germ theory of disease or antibiotics, highlighting the importance of improvements in sanitation and to a lesser extent nutrition in reducing mortality in the nineteenth century (Floud *et al.*, 2011).

Despite these general improvements in health, there is as of yet no consensus on the trajectory of fetal health, or health conditions *in utero*, over the same time period. Woods (2009) and his co-authors (2006) have reconstructed perinatal mortality rates for a number of countries in Western Europe and North America, showing that these rates have declined in the past 150 years. Likewise, several studies have shown that mean birth weights of infants born in hospitals in the nineteenth century were very close to their modern levels (Costa, 1998, 2013; Floud *et al.*, 2011; Ward, 1993). This paper attempts to collate all of this evidence into a coherent story about how fetal health has changed over time.

The paper first defines and discusses the complications in measuring fetal health. It then presents the historical birth weight evidence primarily focussing on the United States with other countries provided as a reference and estimates the influence of changes in environmental and demographic factors on birth weight. It closes with a detailed analysis of trends in perinatal mortality and a discussion of the consequences of the results.

Measuring Fetal Health

Before presenting changes in fetal health in the very long run, it is necessary to discuss what fetal health means and some of the challenges and problems with measuring fetal health in general. Plasticity is very strong in the embryonic and fetal period making the developing child extremely sensitive to changes in conditions *in utero*. Poor conditions such as a nutritional shortage, a lack of key micronutrients, the infection of the placenta or a viral infection can stunt prenatal development harming organ functioning and fetal growth among other negative consequences. Recent research suggests that these conditions and the physiological responses of the fetus to the environment *in utero* may have consequences for the health of an individual across their life course including higher risk of chronic diseases in old age (Godfrey *et al.*, 2007). Thus, the purpose in attempting to measure fetal health over the past 150 years is to understand how the prevalence of unhealthy conditions and unhealthy physiological responses has changed over time and influenced cohort health. This purpose shifts the focus of analysis toward understanding the average health and distribution of health outcomes of the population rather than identifying a subset of individuals that might be at risk. It also leads to an emphasis on conditions that would significantly alter fetal development and the health of surviving infants since these will influence trends in cohort morbidity and mortality.

Given the complexity of prenatal development and the requirements of measuring fetal health, it is very unlikely that any one indicator would be able to perfectly capture fetal health. Thus, we are left with imperfect options from which to choose, especially when pushing measurement into history. Birth weight and length reflect the outcome of fetal growth at one point in time, but they cannot reveal the trajectory of fetal growth before birth. Fetal growth itself is determined by some combination of genetic and epigenetic inheritance as well as dynamic responses to conditions in the womb. Thus, using birth anthropometry, it is impossible to distinguish between an individual born with high inherited growth potential who experiences intrauterine growth restriction and is born at a normal birth weight close to the population mean and an individual of average inherited growth potential who does not experience poor conditions and is born at the same birth weight. Measuring fetal growth directly using ultrasound technology may help ameliorate this problem, but these measurements are not available historically. In addition, fetal growth (and especially weight gain) occurs mostly in the third trimester, so birth weight may not fully capture fetal health in the first and second trimester (Hanson *et al.*, 2015; Roseboom *et al.*, 2011; Wilcox, 2001).

Another potential proxy for fetal health is perinatal mortality since poor conditions *in utero* can lead to stillbirths or early neonatal deaths. Perinatal mortality is especially attractive as a historical proxy since perinatal deaths were systematically registered in a number of countries beginning in the nineteenth century (Woods, 2009). However, perinatal deaths were a relatively rare occurrence even in the nineteenth century when 3-6% of total births ended in a stillbirth or neonatal death. Thus, using perinatal mortality as an indicator of population fetal health could be problematic if the factors that led to these extreme outcomes did not reflect the general, population experience of children during the prenatal period (Wilcox, 2001).

A final indicator of fetal health could be the rate of spontaneous abortions occurring in the population. However, spontaneous abortions are notoriously difficult to measure, and the method employed in the literature of looking at the secondary sex ratio as a proxy for male frailty does not inform about the overall rates of fetal wastage in a population in a way that could be systematically incorporated into the analysis below (Catalano *et al.*, 2008). Thus, this article will focus on birth weights and perinatal mortality.

Birth Weights, 1840 to the present

Rosenberg (1988), Goldin and Margo (1989) and Ward (1993) pioneered the study of historical maternity records containing birth weight that have survived for a number of European and North American maternity hospitals. These hospitals all served slightly different, though mostly working-class populations, in their hinterlands, and they used various selection criteria to admit patients. Supplementary appendix B discusses the inclusion and exclusion criteria for the historical hospitals included in the analysis and discusses selection further. While any given hospital may suffer specific defects, taken together the hospitals can provide a tentative indication of general birth weight levels. As table 1 shows these authors found that birth weight levels in the nineteenth century had already reached modern levels.

Indeed if we compare these with the recent INTERGROWTH-21st standards, nearly all of these populations had birth weights at or above the median birth weight (3,320 grams) for full-term babies (Villar *et al.*, 2014). As Steckel (1998) and Ward (2016) have pointed out, there have been some increases and decreases in birth weights over time. However, these differences are small if the change in birth weight is expressed relative to its standard deviation. Taking the 1985 US population birth weight distribution as a reference (standard deviation of 602 grams), these increases or decreases averaged to 0.18 standard deviations and are all less than 0.45 standard deviations. For reference, to shift the average birth weight in Pakistan in the 1980s (2,770 grams) to the median INTERGROWTH-21st level would require a 0.91 standard deviation increase in birth weight. In addition, the changes in birth weight pale in comparison to the changes in final adult male stature in North America and Europe, which have increased by between 1.4 and 2.3 standard deviations (Hatton and Bray, 2010).

[Table 1 about here]

There is not space in this paper to discuss all of the historical maternity hospitals in detail, but a closer study of the maternity hospitals in Boston, MA may assuage doubts that the high birth weights were driven by the selection of women into each hospital. Ward (1993) collected samples of maternity patient records from three nineteenth-century hospitals in Boston: the New England Hospital for Women and Children (NEH) (1872-1900), the Boston Lying-in Hospital inpatient ward (1886-1900) and the Boston Lying-in Hospital outpatient ward (1884-1900). The NEH and Lying-in inpatient ward provided women a place to give birth and recover afterwards, often for three to four weeks. These hospitals served mainly married and respectable single women. Both required a fee to be paid by the women. The NEH charged $10 per week, which probably excluded some poorer patients from giving birth there. The occupations of women patients in the NEH suggest that most women were from the upper working class or lower middle class. The Lying-in inpatient ward also charged a $20 fee for women resident in Boston, but it did not turn away impoverished women, and 70% of women giving birth in the Lying-in hospital had their fee waived. Thus, the Lying-in inpatient ward seems to have treated a poorer cross-section of the working class along with some complicated pregnancies from higher classes. Both the NEH and Lying-in inpatient ward also served primiparous women, who made up around 60% of births in each hospital (Ward, 1993).

The Lying-in outpatient department was something different altogether. It sent medical students to attend births in women's homes around the city of Boston and then follow up with the women by revisiting them at home for several weeks. The outpatient department was a more equal opportunity operation: it did not charge a fee and it served all women whether respectable or not. Ward (1993) argues that it was the most representative of the working class of the three sets of Boston patient records. It also overwhelmingly served multiparous women.

Given the differences in populations served and selection mechanisms into the three Boston patient record samples, it seems highly unlikely that the high mean birth weights in each could be driven by the hospitals oversampling from the healthy and/or wealthy residents of Boston. What is even more striking is that the birth weight distributions in these hospitals are very similar to the birth weight distribution of the population of white, singleton births, the most comparable group, in Boston in 1985 (Figure 1). The NEH and Lying-in outpatient samples are slightly above the distribution for Boston in 1985 whereas the Lying-in inpatient sample is slightly below. In addition, the birth weight distribution of children born in the Philadelphia Almshouse in the mid nineteenth century is also very similar to the population distribution of white singleton births in Philadelphia in 1985 (see figure A1). Supplementary appendix C carries out statistical tests to determine whether the distributions are statistically different and explains why these tests are generally unhelpful for comparing the historical and modern data.

[Figure 1 about here]

The evidence presented above suggests that mean birth weights and birth weight distributions were very similar in the nineteenth century and today, but what has happened in between? Focussing again on the United States, Costa's (1998) studies of births in New York City in the first half of the twentieth century showed similar levels of birth weights (mean of 3,463 grams). Figure 2 presents the birth weight distributions reported in the Vital Statistics of the United States for every decade from 1950 to 2012. The median birth weight declined by around 20 grams from 1950 to 1970, but the distributions were nearly identical. Between 1970 and 1980 there was a 60 gram increase in the median birth weight and the distribution shifted slightly upward. The median birth weight increased slightly in 1990 and has fallen by 60 grams since then. The birth weight distribution in 2012 has even shifted downward, closely approximating the distribution in 1950.

[Figure 2 about here]

Although less definitive, there is also similar evidence available for Western and Northern Europe. Visser *et al.* (2009) show that birth weight by gestational age percentiles calculated for two Amsterdam clinics between 1931 and 1967 are virtually identical near term to birth weight percentiles calculated from the population of births in the Netherlands in 2001. It is also supported by long run evidence from Norway (Rosenberg, 1988). Thus, despite large improvements in public health and medical technology, the available evidence suggests that birth weight means and distributions have not changed substantially in the past 150 years.

Factors influencing birth weight over time

The remarkable consistency of birth weight means and distributions over time is especially stark considering that many factors influencing birth weight have changed dramatically over the past century. Whereas there has been a general improvement in nutrition and reduction in infection, this article will focus on the influence of maternal height, smoking prevalence among pregnant women, the parity mix of births, and stillbirths and spontaneous abortions because there is more precise information on these to calculate an effect size. Each of these will be discussed in detail in an attempt to understand how each might have affected the birth weight distribution using the United States as a case study.

Maternal height is positively associated with birth weight because the process of maternal constraint prevents a fetus from out-growing the size of the pelvic cavity.

Thus, as women have grown larger over the past 150 years, it is possible that they could have given birth to larger babies. Historical and modern estimates of the influence of maternal height on birth weight controlling for a number of confounding factors are between 9 and 12 grams increase in birth weight per extra centimeter of maternal height (Costa, 1998 and authors' calculations from 2013 US Natality Public Use File). Measuring the secular increase in female height over the past 150 years is somewhat difficult because of the paucity of historical sources, but evidence from nineteenth-century penitentiaries places the mean height of US white women at around 159.5 cm at its lowest point in the nineteenth century (Carson, 2011). This means that the average height of white American women has increased by 3.5 cm to 163 cm for white women giving birth in 2013 (US Natality Public Use File). Thus, if we simply multiplied these two figures, the increase in maternal height in the United States could have accounted for a 31.5 to 42 gram increase in birth weight. However, this likely overestimates the effect because in cross-section 70% of variation in height is driven by genetic or other inherited factors (McEvoy and Visscher, 2009). Thus, part of the relationship between maternal height and birth weight may be reflecting conditions *in utero* but the larger part is reflecting the fact that genetically larger women have larger babies. Thus, the environmental influence would likely be substantially smaller than the genetic influence, diminishing the influence of the secular increase in maternal height on the birth weight distribution. Having said this, the influence of the secular increase in maternal height on the birth weight distribution may be stronger in European countries like the Netherlands that have experienced a 14 cm increase in female average height (de Beer, 2010; Schönbeck *et al.*, 2012).

The rise in tobacco smoking prevalence to the middle of the twentieth century and its subsequent decline could have also influenced the birth weight distribution. Fetal nicotine exposure is associated with intrauterine growth restriction and a number of other pregnancy complications. Studies have also shown that exposure to passive tobacco smoke (second-hand smoke) is nearly as harmful as modest cigarette consumption by the mother, leading to a 200 gram reduction in birth weight. Birth weights are 450 grams lower for mothers who smoke more than ten cigarettes per day (Roquer *et al.*, 1995). Thus, the increase and subsequent decrease in pregnant women's passive and direct exposure to tobacco smoke could have strongly influenced the birth weight distribution. Cohort studies and surveys have generally placed the peak of smoking prevalence for men in the 1960s with the peak for women following a decade or so thereafter (Birkett, 1997; CDC, 2007; Kemm, 2001 and Lund and Lund, 2014). Smoking prevalence has then declined fairly steadily since the 1960s and 1970s with larger declines in North America than in Europe (CDC, 2007; Graham, 1996).

Using Roquer *et al.*'s (1995) estimates of the birth weight penalty for various levels of smoking exposure and the approximate shares of mothers exposed at different points in time, we can estimate how the increase and later decrease in smoking prevalence should have influenced the birth weight distribution in the United States. Table A2 presents these results in detail. Assuming that women did not smoke in the nineteenth century and only received some passive exposure from men (10% of mothers), birth weights were not strongly influenced by nicotine exposure. However, between the nineteenth century and the 1960s and 1970s, nicotine exposure increased dramatically. Assuming that 30% of mothers were never exposed, 30% were exposed passively, 30% were light smokers and 10% were heavy smokers, this increase in nicotine exposure would have decreased the average birth weight by 145 grams. However, the decline in smoking prevalence over the past 40 or 50 years should have increased birth weight counteracting some of this earlier decline. Using the self-reported smoking prevalence in the 2013 US Natality Public Use File covering 87.4% of births in the United States, 5% of mothers were light smokers and 4% were heavy smokers. The US Natality data do not contain information about passive smoke exposure, so 10% exposure was assumed. Taking a new weighted average of Roquer *et al.*'s figures, the decline in smoking over the past 40-50 years would have led to a 118 gram increase in birth weight. Given the large swings in mean birth weight predicted by the changes in smoking prevalence over the past 150 years, the stagnant birth weight distributions are even more striking.

Birth order, or parity, also influences birth weight since first born children tend to have substantially lower birth weights than higher birth order children. Indeed, the difference in birth weight between first born and higher birth order children seems to have been higher in the past, sometimes exceeding 100 grams (see Figures A2 and A3). This could influence the birth weight distribution because the parity mixture of births has changed dramatically over the past century. As total fertility declined from five or six children per woman to less than 2 children per woman, the share of first-born children out of all children born has increased. Figure A4 shows the parity mixture of births in the United States from 1931 to 2000. This does not quite capture the full effect of the fertility decline, which began long before 1931, but unfortunately it is very difficult to estimate the parity mixture of births before national vital registration began in 1931.

In any case, the share of first born children out of all births increased by 17 percentage points between 1931 and the present. It is possible to measure the effect of changes in the parity mix on birth weight by estimating the influence of parity on birth weight in a multivariate regression. Then, one can calculate weighted average birth weights for each time period using the predicted mean values of birth weight for each parity from the regression and the share of births in each parity. Since there appeared to be different historical and modern effects of parity on birth weight, both the predicted average birth weights in the Boston Lying-in outpatient clinic and the U.S. 1985 birth cohort were used. Applying this check to the data, the outpatient clinic birth weights suggest a 25 gram decrease in mean birth weight due to changes in parity mix; however, the decrease is much smaller when using the 1985 birth cohort data at only 10 grams. Thus, it does not appear that changes in the parity mixture of births, at least in the United States since 1931, have led to dramatic changes in the birth weight distribution.

Finally, declining stillbirth rates may have influenced the average health of the surviving population of infants and shifted the birth weight distribution. Before the twentieth century, doctors could do very little to prevent stillbirths and neonatal deaths. However, stillbirth rates and neonatal death rates varied substantially year to year because of viral epidemics, the introduction of lead pipes and associated lead poisoning, venereal diseases, etc., all factors exogenous to the medical treatment of the time. Thus, it is possible to test selection and scarring effects on health from these insults. If the selection effect dominates, then we would expect to see a higher average birth weight when the stillbirth rate was high since these diseases would be more likely to kill foetuses that were already small and unhealthy. On the other hand, if the scarring affect dominated, then we would expect to see lower average birth weights when the stillbirth rate was high because poor intrauterine environments weakened all foetuses, not just those who died. Studies conducted in the past thirty years have mainly found that small to moderate increases in birth weight in general have made up for scientific advancements that have improved survival rates for low birth weight, premature infants (Kramer *et al.*, 2002). However, these studies cover a period where stillbirths and neonatal mortality had already fallen substantially, weakening any selection effects. Thus, it would be helpful to test this for the late nineteenth century.

Figure 3 provides a first, tentative response. It compares the birth weight distributions of babies born in the New England Hospital in Boston between 1872 and 1900 during high stillbirth and low stillbirth years, years where the stillbirth rate in Boston was either above the 75th or below the 25th percentile for the period. The selection effect seems to dominate. The average birth weight is significantly lower in low stillbirth years than in high stillbirth years with the whole distribution of birth weights shifted to the left in low stillbirth years. Thus, declining stillbirth rates or neonatal mortality rates could have led to a decrease in the average birth weight of surviving foetuses, contrary to the expectation that these two indicators should move in opposite directions. Evidence on spontaneous abortions also suggests that the selection effect dominated. In a study of nineteenth century Finland, Bruckner *et al.* (2014) found that males in cohorts with lower secondary sex ratios, when male culling was higher, were healthier because they were less likely to die as infants. If the selection effect of spontaneous abortion and stillbirths dominated, it raises significant questions about whether these indicators can be used as a proxy for fetal health at all. A strong selection effect suggests that perinatal mortality is more a proxy of poor health among the foetuses and infants that die, not poor health across the whole population.

[Figure 3 about here]

To summarize, there have been a number of factors influencing the birth weight distribution over time. Although the estimations presented above were conducted for the United States, these effects would be similar in other countries even if the turning points of the trends and the absolute magnitude of the effects might be somewhat different. The secular increase in maternal height and the decline in nicotine exposure since the 1960s would have shifted the birth weight distribution upward whereas the increase in smoking prevalence before the 1960s, the shift toward lower parity births, and the decline in the stillbirth rate may have shifted the birth weight distribution downward. The fall in median US birth weight over the past twenty years may have been driven by the rise of induced labor and fraternal twins conceived through *in vitro* fertilization (Zhang *et al.*, 2010). However, the net effect of all of these forces seems to have been a static birth weight distribution over the past 150 years. Thus, the important research question, beyond the scope of this paper, is whether the short and long-run health benefits or consequences of these various shifts are equal. Is birth weight perfectly able to proxy the health costs and benefits of these various exposures, with static birth weight distributions suggesting that there simply has not been improvement in fetal health over the past 150 years? Or for instance, is the health cost of nicotine exposure *in utero* more severe than the health benefit from having a taller, healthier mother? Using birth weight as a simple proxy for fetal health does not enable us to distinguish between these underlying mechanisms and limits our ability to determine whether clinical and policy interventions have improved or will improve health across the life course.

Causes of Decline in Perinatal Mortality

Although birth weight distributions seem to have been relatively constant over the past century and a half, there have been substantial declines in perinatal mortality during the same time period. The selection effect of stillbirths on birth weight in late nineteenth-century Boston has already raised doubts about whether perinatal mortality accurately proxies the health *in utero* of the general population of births, but it is possible that substantial improvements in the perinatal mortality rates (rather than fluctuations around a stationary trend) could have marked an improvement in fetal health. Perinatal mortality rates were also recorded more systematically and widely in historical periods than birth weights, making them a potentially attractive proxy for fetal health.

However, stillbirth (SBR) and early neonatal mortality rates (ENMR) are not without their problems. Most importantly from the perspective of measuring fetal health, stillbirths and early neonatal deaths may occur because of infection or intra-uterine growth restriction, factors that indicate poor fetal health, or because of intrapartum complications, which could be unrelated to fetal health especially in a period before the widespread use of Caesarean sections. Thus, it will be important to understand why perinatal mortality declined in order to interpret whether this decline suggests improvement in fetal health or not. In addition, SBRs and ENMRs suffer inconsistencies in registration across countries from differing definitions of stillbirths and cultural practices around baptism of infants. Registration appears to have been fairly good in Scandinavian countries but more problematic in Catholic countries such as France and Italy (Woods, 2009). Another potential problem arises from the periods for which perinatal mortality data are available. While stillbirth rates were recorded for Scandinavian countries since the mid eighteenth and early nineteenth centuries, these data were not systematically registered in the US, the UK, France and Italy until well into the twentieth century. Woods (2009) analysed historical trends in stillbirths across a wide range of countries accounting for the various registration and other problems with the data. His analysis will form the basis of the discussion below.

Figures 4 and 5 show the trends in SBR and ENMR for countries in Europe and North America from the 1880s to the present. The trends, if not the starting levels, are very similar. SBRs and ENMRs were relatively stagnant from the 1880s until the late 1930s when they began to decline dramatically. Woods (2009) attributed the differences in level between countries in the early twentieth century to obstetric practice in each country. Scandinavian countries and the Netherlands had larger numbers of highly trained doctors and midwives who promoted best practice, and therefore, they had substantially lower stillbirth rates than England, the United States, France or Italy (Woods *et al.*, 2006). However, all of these countries experienced a sharp decline in SBR and ENMR in the late 1930s. The common timing of the decline is extraordinary and rules out any country-specific environmental, registration or health policy changes. Thus, Woods (2009) suggested that the sharp decline could have been caused by the introduction of sulfa drugs and antibiotics from the mid 1930s onward. However, he was not able to determine the precise mechanism through which the adoption of antibiotics would have influenced the SBR since bacterial infections were not a leading cause of stillbirths.

[Figures 4 and 5 about here]

Løkke (2012) has provided the most extensive evidence to date linking the decline in perinatal mortality to the introduction of antibiotics. Studying reports on causes of death from the National Hospital in Copenhagen, Denmark between 1910 and 1975, she found that antibiotics reduced perinatal mortality through two mechanisms. First, antibiotics allowed physicians to conduct more invasive surgeries to protect the life of the fetus than they had in the past. In the pre-antibiotic era, Danish obstetricians placed the highest value on preventing maternal mortality, so they only carried out Caesarean sections or other invasive surgeries when the mother's life was at risk. However, once Prontosil and Sulfa drugs became available in the mid-1930s followed by Penicillin in the mid-1940s, Danish obstetricians began performing more invasive intrapartum surgeries because they could protect mothers from puerperal fever. The most common intrapartum complications affected were placental abruption, placenta praevia, eclampsia, contracted pelvis and a prolapsed umbilical cord. Løkke found that reduced mortality of mothers and infants treated for these five complications in the National Hospital accounted for over half of the decline in perinatal mortality in Copenhagen between 1937 and 1957. Thus, the majority of the decline in perinatal mortality appears to have been driven by improvements in intrapartum care rather than change in health conditions *in utero*.

However, Løkke's second mechanism does allow for improvements in fetal health. She argues that antibiotics accelerated the cohort improvement in women's health across the twentieth century because they reduced infections during pregnancy. A general reduction in infections during pregnancy likely improved maternal health and thus fetal health on its own, but most important from the perspective of the decline in perinatal mortality was a reduction in women entering the hospital with syphilis. 5.8% of mothers in 1927 entered the hospital with syphilis, but this rate had fallen to 0.1% by 1957. This reduction in the syphilis rate may explain the decline in perinatal mortality not attributed to the improvements in intrapartum care described above.

Thus, if Denmark is representative, it appears that much of the decrease in perinatal mortality beginning in the late 1930s can be attributed to better interventions in response to intrapartum complications and to the reduction in syphilis prevalence made possible by antibiotics. However, neither of these factors accurately reflect the health of the population in general. The vast majority of children did not experience serious intrapartum complications and syphilis only affected those infected. Thus, most of the decline in perinatal mortality was not caused by a general improvement in fetal health. Rather, it was made possible by preventing deaths among foetuses and neonates most at risk. This conclusion is bolstered by evidence that even since the mid-1970s in the United States, reductions in perinatal mortality have been mainly driven by better neonatal care for at-risk infants and that most nutritional interventions have failed to reduce intrauterine growth restriction or preterm births (Goldenberg and Culhane, 2007). These findings, along with the selection effects of stillbirths shown for late nineteenth-century Boston above, raise serious questions about whether declines in perinatal mortality reflect improvements in fetal health over the past 150 years.

Discussion

The available evidence seems to suggest that fetal health has not improved substantially over the past 150 years in the United States or in Northern and Western Europe though the data is more limited. Birth weight averages and distributions, despite minor changes, have remained remarkably similar over time and the decline in perinatal mortality beginning in the late 1930s was driven by factors unrelated to the general health *in utero* of the population. This striking and somewhat shocking finding could point toward two conclusions, both of which have critical implications for contemporary biomedical and social science research. On the one hand, it is possible that fetal health has remained stagnant over the past 150 years despite substantial improvements in living standards, medical technology and health infrastructure. Drastic improvements in nearly every aspect of postnatal health perhaps did not strongly influence prenatal health because foetuses were taking all of the resources they needed from their mothers before the improvements in health. If this is true, then improving fetal health cannot explain cohort improvements in life expectancy and mortality risk age profiles or gains in human capital over the twentieth century (Almond and Currie, 2011; Arora, 2013). These cohort effects may have been driven by the reduction in infant and childhood scarring from diseases and undernutrition (Crimmins and Finch, 2006; Hatton, 2014; Quaranta, 2014).

On the other hand, the evidence may suggest that the indicators used to measure fetal health, indicators that are still widely used today, are not as helpful as researchers might hope. Perhaps fetal health has improved over the past 150 years, but we do not currently have indicators that can capture this improvement. For instance, micronutrient supplementation (iodine and folic acid) along with a reduction in rubella infections among pregnant women after the introduction of the MMR vaccination must have led to an improved level of fetal health even if there were no shift in the birth weight distribution. This would suggest looking for new ways to measure fetal health in the past and future. Birth weight is an especially problematic indicator since it proxies fetal health with so much error. As mentioned above, it provides little information about the growth trajectory *in utero*. In addition, birth weight is most sensitive to poor conditions in the third trimester, but may be unaffected by health shocks earlier in the prenatal period (Roseboom *et al.*, 2011). Thus, from a statistical perspective, regressions using birth weight to predict later life outcomes could suffer from omitted variable bias and/or attenuation bias, seriously compromising the parameter estimates and any causal inference about the relationships tested.

In addition, these historical findings question the standards for birth weight by gestational age developed recently by the INTERGROWTH-21st group based on the pregnancies of healthy, non-obese women from eight different countries (Villar *et al.*, 2014). If birth weights have remained remarkably stable in the United States and Western and Northern Europe over a period where health improved so dramatically, can we expect to see convergence to the 'optimal' birth weight means and distributions as developing countries experience a similar transition? If not, then why do birth weights vary so dramatically around the world? This historical evidence suggests that it is difficult to define any single pattern of fetal growth, even one derived as precisely as Villar *et al.*'s pattern, as optimal (Hanson *et al.*, 2015).

In order to move past these problems, researchers need to focus on developmental pathways that might influence health rather than on one simple indicator, such as birth weight. For social scientists this might involve studying factors influencing fetal health directly rather than relying on birth weight to capture the net effects. This is especially important since there may be large degrees of confounding between inherited and socioeconomic factors that could influence both birth weight and later health outcomes. Researchers could study how maternal smoking during pregnancy affects the later life health of a fetus. In addition, it would be interesting to know whether demographic factors such as parity, mother's age and gestational age that influence birth weight would also produce poor later life health outcomes (c.f. Gavrilov and Gavrilova, 1997). Researchers could also incorporate a wider range of proxies for fetal health such as the Ponderal index, birth length, head circumference and placental weight in an attempt to better proxy the fetal environment. These indicators are even available in some nineteenth-century maternity hospital records as well, though this information would need to be collected from original sources in the archive. Medical researchers could focus on epigenetic markers, which could be used as additional proxies for development trajectories *in utero*. Thus, only by studying the developmental process directly will we be able to move beyond birth weight to a greater and more complex understanding of fetal health.

References

Almond, D., & Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives*, *25*(3), 153–172. http://doi.org/10.1257/jep.25.3.153

Arora, S. (2013). Understanding Aging During the Epidemiologic Transition. *Research in Economic History*, *29*, 1–69. http://doi.org/10.1108/S0363-3268(2013)0000029004

Barker, D. (1997). Maternal nutrition, fetal nutrition, and disease in later life. *Nutrition*, *13*(9), 807–813.

de Beer, H. (2010). Physical stature and biological living standards of girls and young women in the Netherlands, born between 1815 and 1865. *The History of the Family*, *15*(1), 60–75. http://doi.org/10.1016/j.hisfam.2009.12.003

Birkett, N. J. (1997). Trends in smoking by birth cohort for births between 1940 and 1975: a reconstructed cohort analysis of the 1990 Ontario Health Survey. *Preventive Medicine*, *26*(4), 534–541. http://doi.org/10.1006/pmed.1997.0169

Bruckner, T. A., Helle, S., Bolund, E., & Lummaa, V. (2014). Culled males, infant mortality and reproductive success in a pre-industrial Finnish population. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1799), 20140835. http://doi.org/10.1098/rspb.2014.0835

Carson, S. A. (2011). Height of female Americans in the 19th century and the antebellum puzzle. *Economics and Human Biology*, *9*(2), 57-164.

Catalano, R., Bruckner, T., & Smith, K. R. (2008). Ambient temperature predicts sex ratios and male longevity. *Proceedings of the National Academy of Sciences*, *105*(6), 2244–2247. http://doi.org/10.1073/pnas.0710711104

Center for Disease Control and Prevention. (2007). Cigarette smoking among adults – United States, 2006. *Morbidity and Mortality Weekly Report*, *56*(4), 1157-1161.

Conley, D., Strully, K. W., & Bennett, N. G. (2003). *The Starting Gate: Birth Weight and Life Chances*. Berkeley: University of California Press.

Costa, D. L. (1998). Unequal at birth: A long-term comparison of income and birth weight. *Journal of Economic History*, *58*, 987–1009.

Costa, D. L. (2013). Health and the Economy in the United States, from 1750 to the present. *NBER Working Paper*, no. 19685.

Crimmins, E., & Finch, C. (2006). Infection, inflammation, height, and longevity. *Proceedings of the National Academy of Sciences*, *103*(2), 498–503.

Figlio, D., Guryan, J., Karbownik, K., & Roth, A. J. (2014). The Effects of Poor Neonatal Health on Children's Cognitive Development. *American Economic Review*, *104*(12), 3921–3955. http://doi.org/10.1257/aer.104.12

Floud, R., Fogel, R. W., Harris, B., & Hong, S. K. (2011). *The Changing Body: Health, Nutrition and Human Development in the Western World since 1700*. Cambridge: Cambridge University Press.

Godfrey, K. M., Lillycrop, K. A., Burdge, G. C., Gluckman, P. D., & Hanson, M. A. (2007). Epigenetic Mechanisms and the Mismatch Concept of the Developmental Origins of Health and Disease. *Pediatric Research*, *61*(Supplement), 5R–10R. http://doi.org/10.1203/pdr.0b013e318045bedb

Goldenberg, R. L., & Culhane, J. F. (2007). Low birth weight in the United States. *American Journal of Clinical Nutrition*, *85*(2), 584S–590S.

Goldin, C., & Margo, R. A. (1989). The poor at birth: Birth weights and infant mortality at Philadelphia's Almshouse Hospital, 1848–1873. *Explorations in Economic History*, *26*(3), 360–379.

Graham, H. (1996). Smoking prevalence among women in the European Community 1950–1990. *Social Science & Medicine*, *43*(2), 243–254.

Hanson, M., Kiserud, T., Visser, G. H. A., Brocklehurst, P., & Schneider, E. (2015). Optimal Fetal Growth: A Misconception? *American Journal of Obstetrics and Gynecology*, *213*, 332-334. http://doi.org/10.1016/j.ajog.2015.06.027

Hatton, T. J. (2014). How have Europeans grown so tall? *Oxford Economic Papers*, *66*(2), 349–372. http://doi.org/10.1093/oep/gpt030

Hatton, T. J., & Bray, B. E. (2010). Long run trends in the heights of European men, 19th-20th centuries. *Economics and Human Biology*, *8*(3), 405–413. http://doi.org/10.1016/j.ehb.2010.03.001

Kemm, J. R. (2001). A birth cohort analysis of smoking by adults in Great Britain 1974-1998. *Journal of Public Health Medicine*, *23*(4), 306–311.

Kramer, M. S., Morin, I., Yang, H., Platt, R. W., Usher, R., McNamara, H., et al. (2002). Why are babies getting bigger? Temporal trends in fetal growth and its determinants. *The Journal of Pediatrics*, *141*(4), 538–542. http://doi.org/10.1067/mpd.2002.128029

Løkke, A. (2012). The antibiotic transformation of Danish obstetrics. The hidden links between the decline in perinatal mortality and maternal mortality in the mid-twentieth century. *Annales de démographie historique*, *123*, 205-224.

Lund, I., & Lund, K. E. (2014). Lifetime smoking habits among Norwegian men and women born between 1890 and 1994: a cohort analysis using cross-sectional data. *British Medical Journal Open*, *4*, 1–9. http://doi.org/10.1136/bmjopen-2014-005539

McEvoy, B., & Visscher, P. (2009). Genetics of human height. *Economics and Human Biology*, *7*(3), 294-306.

Quaranta, L. (2014). Early life effects across the life course: The impact of individually defined exogenous measures of disease exposure on mortality by sex in 19th- and 20th-century southern Sweden. *Social Science & Medicine, 119*, 266-273.

Roquer, J. M., Figueras, J., Botet, F., & Jiménez, R. (1995). Influence on fetal growth of exposure to tobacco smoke during pregnancy. *Acta Paediatrica*, *84*(2), 118–121. http://doi.org/10.1111/j.1651-2227.1995.tb13592.x

Roseboom, T. J., Painter, R. C., van Abeelen, A. F. M., Veenendaal, M. V. E., & de Rooij, S. R. (2011). Hungry in the womb: What are the consequences?

Lessons from the Dutch famine. *Maturitas*, *70*(2), 141–145. http://doi.org/10.1016/j.maturitas.2011.06.017

Rosenberg, M. (1988). Birth weights in three Norwegian cities, 1860-1984. Secular trends and influencing factors. *Annals of Human Biology*, *15*(4), 275–288.

Schönbeck, Y., Talma, H., van Dommelen, P., Bakker, B., Buitendijk, S. E., HiraSing, R. A., & van Buuren, S. (2012). The world's tallest nation has stopped growing taller: the height of Dutch children from 1955 to 2009. *Pediatric Research*, *73*(3), 371–377. http://doi.org/10.1038/pr.2012.189

Steckel RH. (1998). Birth weights and stillbirths in historical perspective. *European Journal of Clinical Nutrition*. *52*(Suppl 1), S16-20.

Villar, José *et al.* (2014). International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *Lancet*, *384*, 857-868.

Visser, G. H. A., Eilers, P. H. C., Elferink-Stinkens, P. M., Merkus, H. M. W. M., & Wit, J. M. (2009). Early Human Development. *Early Human Development*, *85*(12), 737–744. http://doi.org/10.1016/j.earlhumdev.2009.09.008

Ward, W. P. (1993). *Birth Weight and Economic Growth: Women's Living Standards in the Industrializing West*. Chicago: University of Chicago Press.

Ward, W. P. (2016). Birth Weight as an Indicator of Human Welfare. (J. Komlos & I. R. Kelly, Eds.) (Vol. 1, pp. 621-631). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199389292.013.33

Wilcox, A. J. (2001). On the importance—and the unimportance— of birthweight. *International Journal of Epidemiology*, *30*(6), 1233–1241. http://doi.org/10.1093/ije/30.6.1233

Woods, R. I., Lokke, A., & van Poppel, F. (2006). Two hundred years of evidence-based perinatal care: late-fetal mortality in the past. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *91*(6), F445–F447. http://doi.org/10.1136/adc.2006.098178

Woods, R. I. (2009). *Death before birth: Fetal health and mortality in historical perspective*. Oxford: Oxford University Press.

Zhang, X., Joseph, K. S., & Kramer, M. S. (2010). Decreased term and postterm birthweightin the United States: impact of labor induction. *American Journal of Obstetrics and Gynecology*, *203*(2), 124.e1–124.e7. http://doi.org/10.1016/j.ajog.2010.03.044

Main Text Figures and Tables

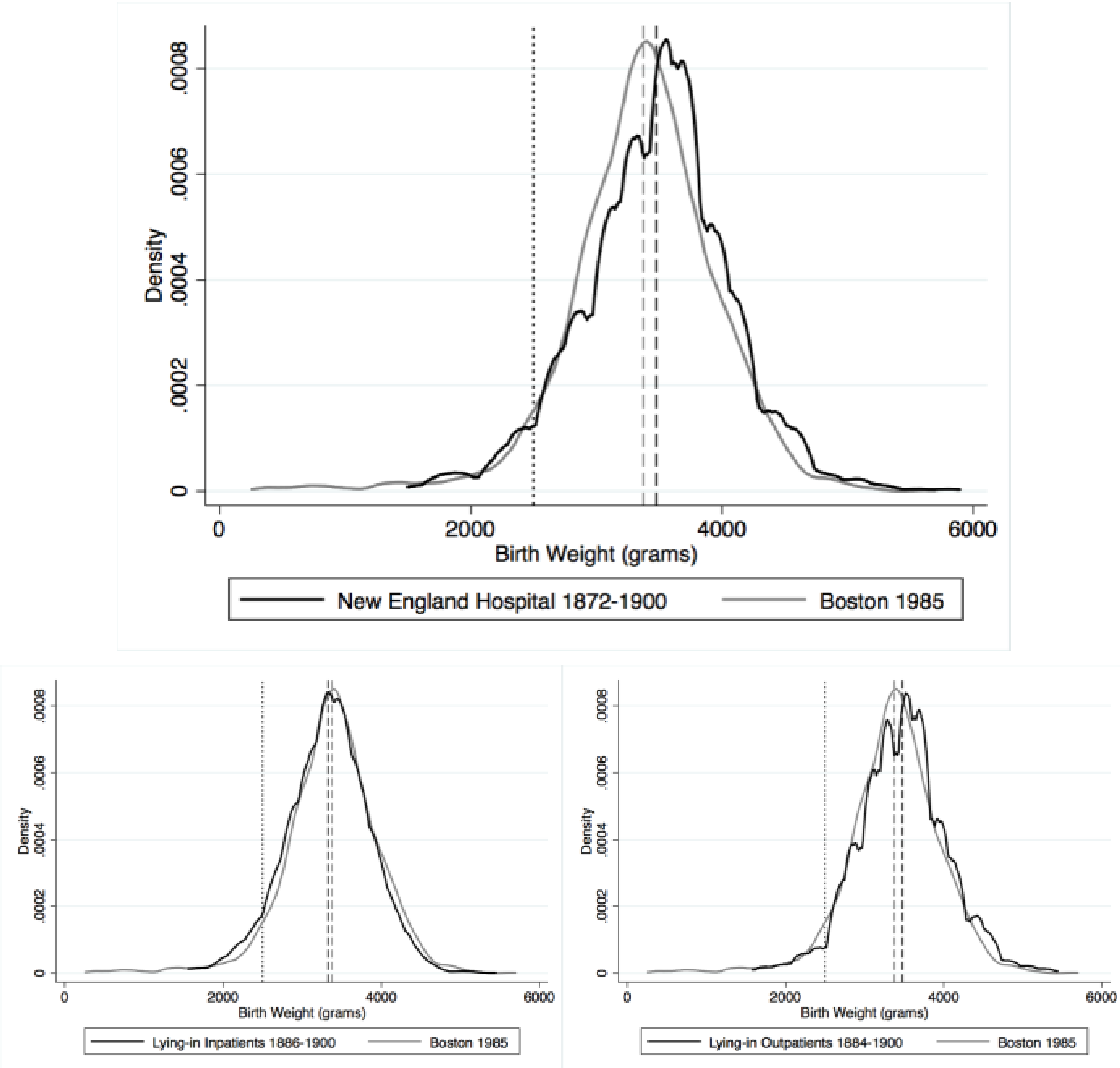
Table 1: Birth weights of children in historical maternity hospitals compared with children of like populations born in the 1970s and 1980s.

	Historical			Modern (1970s-80s)		
City	Years	Mean Birth Weight (g)	% Low Birth Weight	Year	Mean Birth Weight (g)	% Low Birth Weight
North America						
Boston	1872-1900 (NEH)	3,480	5.8%			
Boston	1886-1900 (In)	3,330	6.5%	1985	3,377	5.6%
Boston	1884-1900 (Out)	3,479	4.3%			
Philadelphia	1848-1873	3,406	7.5%	1985	3,373	5.5%
Montreal	1847-1899	3,403	6.5%	1988	3,303	6.3%
Europe						
Dublin	1869-1899	3,203	10.8%	1978-79	3,473	4.4%
Edinburgh	1847-1899	3,227	11.1%	1985	3,342	6.6%
Norway	1860-1900	3,400-3,500				
Vienna	1865-1899	3,098	9.3%	1978	3,320	5.8%

Notes: See supplementary appendix B for more details on the characteristics of each maternity hospital. LBW means low birth weight, a birth weight under 2,500 grams.

Sources: see Table B2 for historical sources. Modern data: Boston and Philadelphia – white, singleton births from the U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990); Montreal, Dublin and Vienna – Ward (1993, p. 134); Edinburgh – Bonellie *et al.* (2008); National Statistics for Scotland (2014).

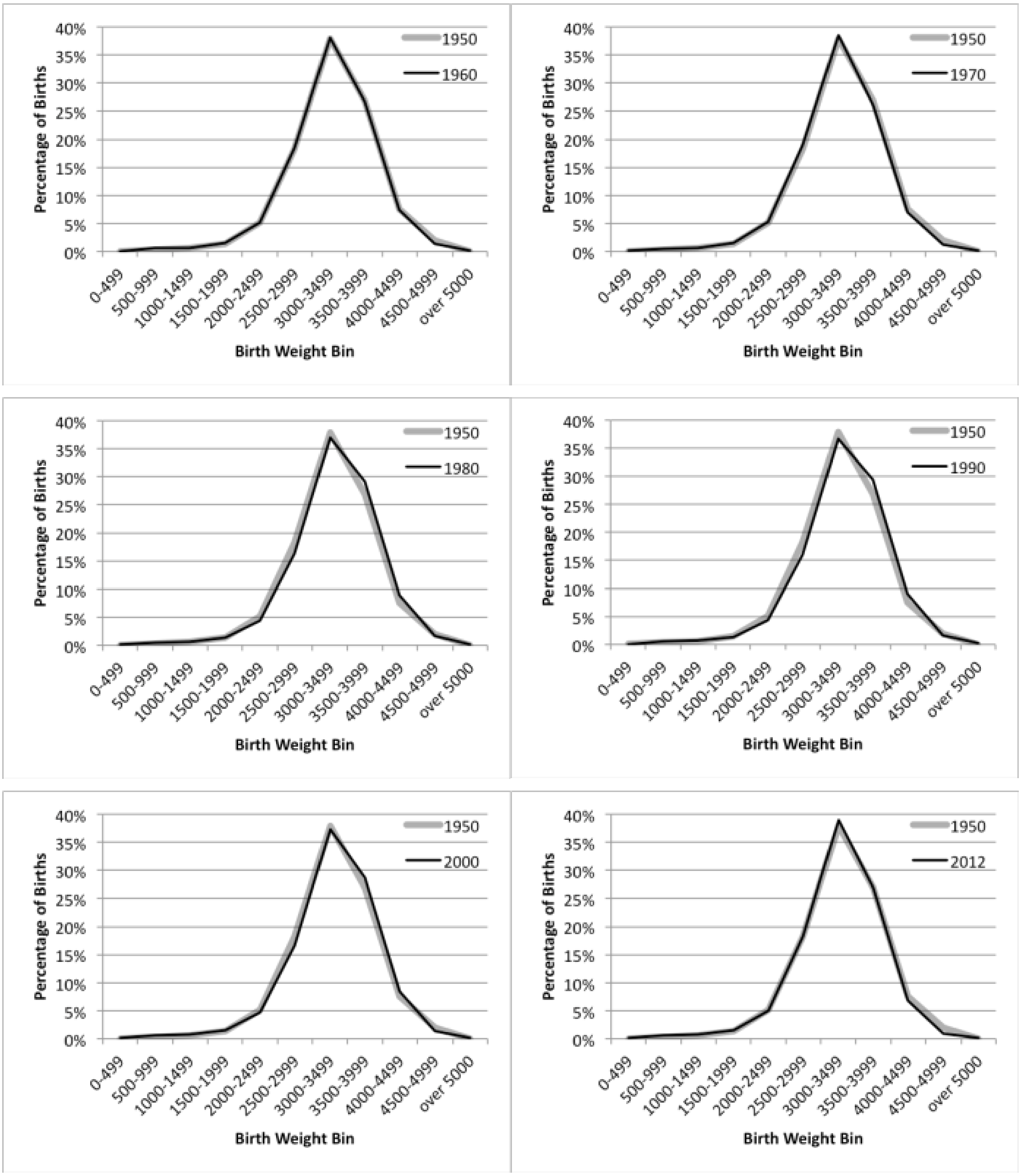
Figure 1: Birth weight distributions in historical hospitals in nineteenth-century Boston compared to the population distribution of Boston in 1985.



Notes: The dotted vertical line marks the low birth weight cut-off of 2,500 grams. The dashed vertical lines mark the mean of each distribution. The distribution of birth weights in Boston in 1985 only includes white, singleton births to make it most comparable with the historical data.

Sources: Historical birth weight datasets – Gagné and Ward (2012); Birth weights in Boston 1985 – U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990). See supplementary appendix D for references.

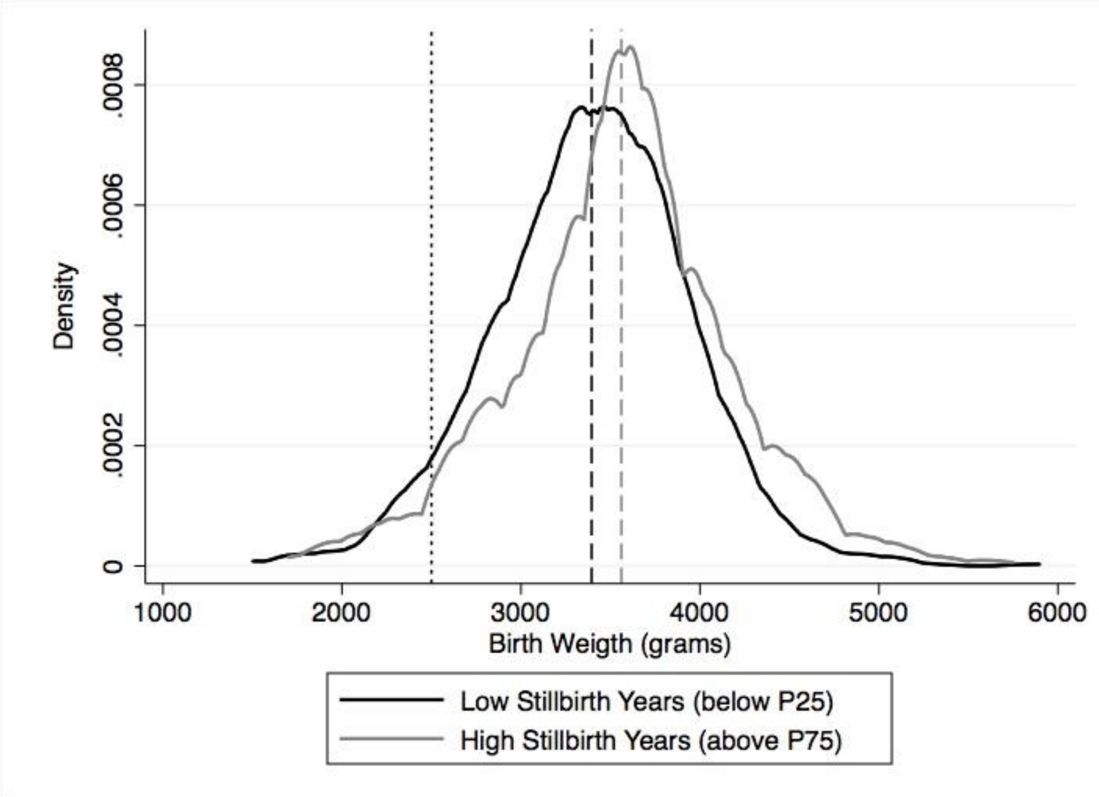
Figure 2: Stagnation of birth weight distributions in the United States from 1950 to the present.



Notes: Distributions before 1990 are based on a 50% sample of registered births in the 50 states of the United States. From 1990 onward they are based on the population of births.

Sources: Vital Statistics of the United States, 1950-2012. See supplementary appendix D for references.

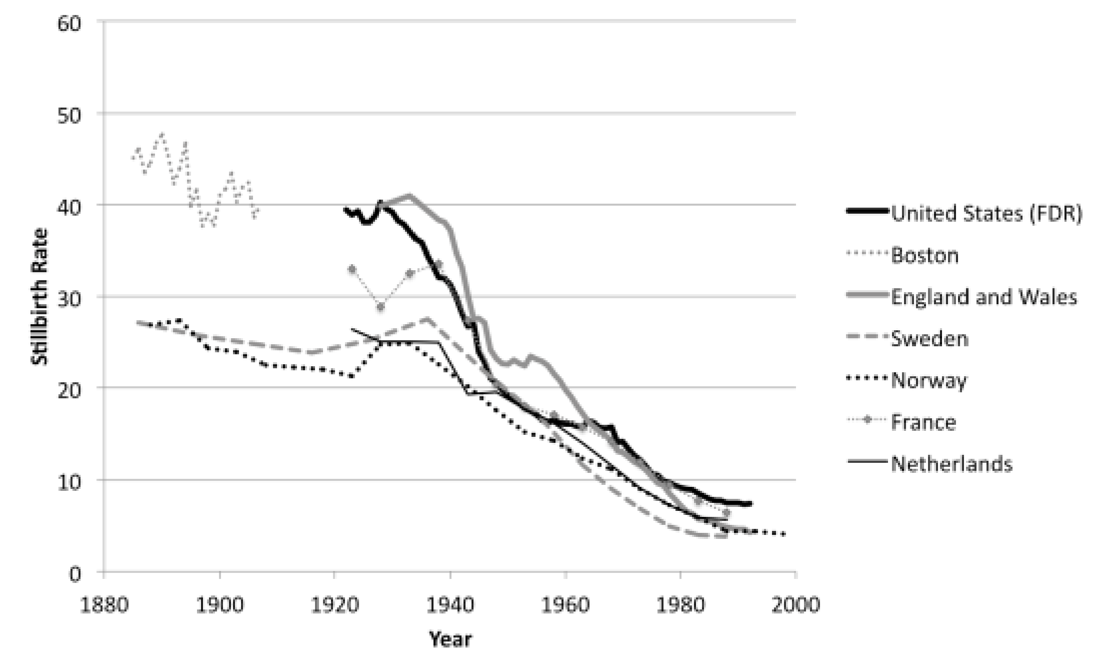
Figure 3: Birth weight distributions of children born in the New England Hospital, Boston in high and low stillbirth rate years, 1872-1900.



Notes: The average stillbirth rate in high and low stillbirth rate years were 51.06 and 39.96, respectively.

Sources: Gagné and Ward (2012); Registry Department, City of Boston (1908, pp. 306-307). See supplementary appendix D for references.

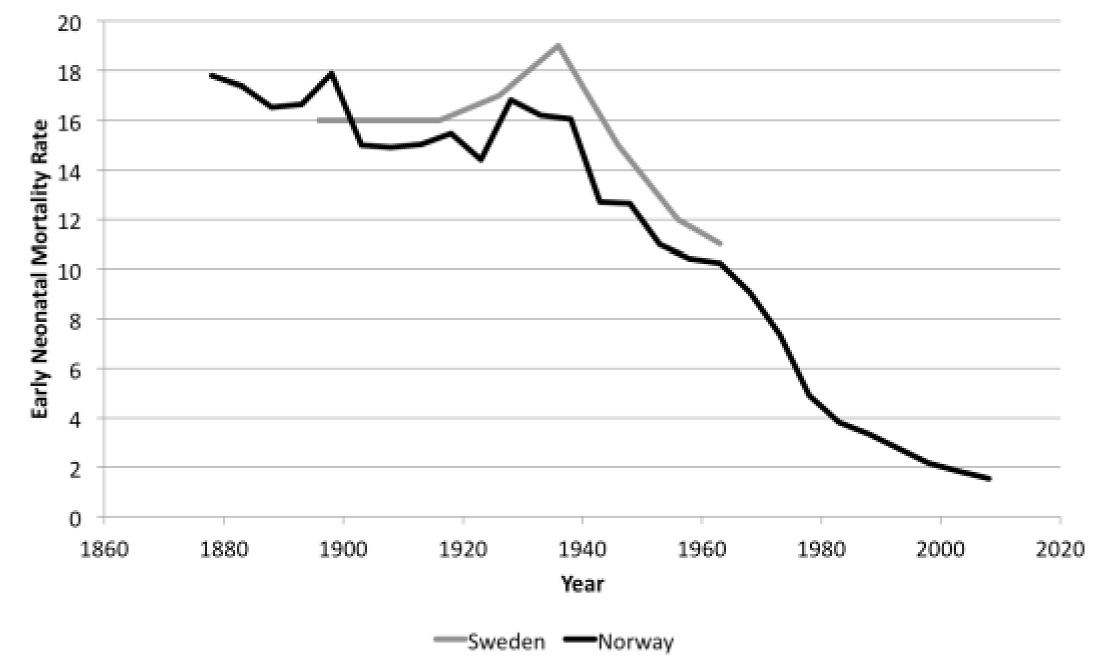
Figure 4: Stillbirth rates (SBR) in Western Europe and North America declined simultaneously beginning in the late 1930s.



Notes: Data for the United States are fetal death ratios (FDR), stillbirths over live births, rather than stillbirth rates, stillbirths over total births. Data for Norway, the Netherlands and France are quinquennial averages, and data for Sweden are decadal averages until 1961 and quinquennial averages thereafter. Data before 1921 for France and the Netherlands are not reported because an unknown number of early neonatal deaths were recorded as stillbirths.

Sources: United States – Haines (2006, table Ab912-927); Boston – Registry Department, City of Boston (1908, pp. 306-307); England and Wales – Office of National Statistics (2014); Sweden – *Historisk Statistik för Sverige* (1967, pp. 108-109) and Macfarlane *et al.* (2000, pp. 664-665); Norway – Statistics Norway (2014, table 05860); France and the Netherlands – Macfarlane *et al.* (2000, pp. 664-665). See supplementary appendix D for references.

Figure 5: Early neonatal mortality rates (ENMR) declined sharply at the end of the 1930s.



Notes: Data for Norway are quinquennial averages, and data for Sweden are decadal averages until 1961 and quinquennial averages thereafter.

Sources: Sweden – *Historisk Statistik för Sverige* (1967, pp. 115-116); Norway – Statistics Norway (2014, table 05860). See supplementary appendix D for references.

Web Appendix A: Supplementary Figures and Tables

[Table A1 here]

[Figure A1 here]

[Table A2 here]

[Figures A2, A3 and A4 here]

Web Appendix B: Characteristics of Historical Maternity Hospitals

As mentioned in the article text, the historical maternity hospitals did not draw a random sample of women giving birth in their catchment areas. Instead, they all targeted (selected) different subsections of the population. Likewise, the policies of the hospitals changed over time, especially after the turn of the twentieth century when larger numbers of women began giving birth in hospital. In addition, not all of the records were as carefully kept as others. Thus, it is necessary to set some inclusion and exclusion criteria for this meta-study to ensure that poor data quality and selection effects do not skew the analysis. Inclusion criteria will be discussed first before turning to exclusion criteria.

The first inclusion criterion was that the hospital needed to record birth weights for the nineteenth century. Focusing on the nineteenth century is necessary because in the early twentieth century, maternity hospitals began to serve larger and larger portions of the population rising from percentages around 5-10% to more than 50% of births. Thus, it is extremely difficult to determine whether changes in mean birth weight or the low birth weight percentage mark real changes in population health or are simply driven by changes in the selection of people into the hospital. Limiting the calculations to the nineteenth century, even if the hospital records extended to the 1930s, ensures that these major changes in selection into hospitals do not strongly influence the results (more on this below).

The hospital records also needed a large sample size and a few key covariates in order to understand the differences in the hospital. Fortunately, sample size was not an issue since all of the datasets had at least 2,000 observations. They also had the key covariates necessary to determine selection and potentially run regressions to control for these factors: parity, mother's age, marital status and sex of the child. Based on these inclusion criteria, there were eleven potential maternity hospital datasets that could be included in the study. The basic statistics for each dataset are presented in Table B1.

[Table B1 about here]

Moving to exclusion criteria, selection and representativeness of the patients are foremost. Table B2 lists a number of ways in which the hospital samples may have been selected. The hospitals were responsible for different shares of total births in their local area. They varied in the class groups that they targeted either directly or indirectly. Some specialised in helping primiparous, normally unmarried women whereas others had a much stronger focus on married and multiparous women. Very few charged a fee to their patients and several that did charge a fee waived the fee in the case of destitute patients. Some used nineteenth-century morals of respectability to select the women they would help whereas others would happily care for prostitutes. Some of these criteria likely affect the mean birth weight in each city. The fact that nearly all of the hospitals served the poor and working classes suggests that the birth weights will tend to be lower than the actual population figures. In addition, hospitals that specialised in serving primiparous women will likely have lower mean birth weights than the population since first-born children tended to weigh at least 100 gram less than their higher birth order siblings. The fees and concerns about morality might lead to an overstatement of mean birth weight if the hospitals excluded the poorest who might not have been able to pay or did not fit into middle-class conceptions of respectability.

[Table B2 about here]

However, none of these criteria merited exclusion from the analysis. This is in part because the paucity of historical data prevents one from being overly picky, but it is also because the selection along these mechanisms was mostly able to be observed by viewing changes in the covariates. More worrying is the indication that two of the datasets had a large degree of unobservable selection directly related to health outcomes. Each of these hospitals will be discussed at length to justify their exclusion from the analysis conducted above.

The first excluded hospital was the *Ospedal Sant'Orsola* in Bologna (Ward, 2004). Mean birth weights increased in the *Ospedal* over time from below 3,000 grams in the 1880s to above 3,300 grams in the 1930s. Over the same period the low birth weight percentage declined from nearly 19% to 6%. This trend could suggest major improvements in maternal and fetal health in Bologna over this period, but there is evidence that the hospital may have been selecting individuals with poor health in the earlier period. The stillbirth rate in the *Ospedal* was twice as high in the 1880s and 1890s as the Bologna city level and then converged to the city level over time. This evidence suggests that the hospital may have been serving mainly problem births in the nineteenth century, explaining at least part of the decrease in the low birth weight percentage and increase in mean birth weight over time. This evidence of direct selection on health requires that the *Ospedal* be excluded from the study.

The second hospital to be excluded is the *Algemeen Ziekenhuis* in Utrecht (Ward, 2003). This was one of two hospitals that Ward analysed in Utrecht, and unfortunately it is the only hospital for which data has been deposited online. The *Algenmeen Ziekenhuis* went from delivering 4% of births in Utrecht in the nineteenth century to 17% in the 1930s, but it was overshadowed by the *Polikliniek*, which opened in the early 1900s. The *Polikliniek* conducted 40% of births in Utrecht within a few years of opening, and this percentage increased to 75% by 1940 (Ward, 2003, p. 383). In Ward's discussion of the *Algemeen Ziekenhuis*, he states that 'women with problematic pregnancies, difficulties in labour, or post-delivery medical need often turned to the hospital' (Ward, 2003, p. 385). This statement is corroborated by the fact that when the *Polikliniek* and *Algemeen Ziekenhuis* first overlap in the early twentieth century, the difference in the mean hospital birth weights was nearly 300 grams and the difference in the low birth weight percentage was 10 percentage points. Thus, it seems clear that the *Algemeen Ziekenhuis* selected patients on the basis of their poor health and must be excluded from the analysis.

The application of the inclusion and exclusion criteria leaves nine historical maternity hospitals which can be analyzed. This is obviously not a huge amount of evidence to draw upon, but given the paucity of historical information on birth weight, it is the widest range of evidence for North America and Northern and Western Europe currently available.

Web Appendix C: Statistical Tests on Birth Weight Distributions

Looking at the historical and modern birth weight distributions presented in Figures 1 and 2, readers may wonder whether the differences in the means and distributions between the historical samples and modern population are statistically significant. This is a valid question, so this appendix provides details on the statistical results. However, in general these tests are not very helpful for understanding changes in mean birth weights and birth weight distributions over time for three reasons. First, when comparing populations, the meaning of inferential statistics is dubious. Second, large sample sizes can make differences in the means and distributions statistically significant, but that does not mean that the actual magnitude of the differences are meaningful. And third, non-parametric tests that compare distributions are very sensitive to levels of rounding and heaping in the data, which are present in the historical birth weight datasets. Each of these points will be discussed in turn before presenting some of the statistical results below.

The first reason that statistical tests are not particularly helpful is that their meaning is unclear when comparing populations as in Figure 2. When using inferential statistics, one usually is working with a sample drawn from a population, and the standard error tells you the standard deviation of the sampling distribution of whatever sample statistic you are estimating. The confidence intervals calculated for the statistic (based on the standard error) give you a range of values in which you can have confidence that the population statistic will lie. Thus, if the mean birth weight of a sample of babies is 3,350 grams plus or minus 50 grams for a 95% confidence interval, then there is a 95% probability that the population mean lies between 3,300 and 3,400 grams. However, in the case of the distributions compared in Figure 2, the population means are known. Thus, it is unclear what statistical tests will tell us about the differences that cannot be learned simply by comparing the means and deciding whether the differences are meaningful or not.

The second reason that statistical tests can be misleading when comparing birth weight distributions is that the sample sizes can be extremely large. The standard error is strongly influenced by sample size, so the larger the sample size the lower the standard errors will be and the more likely one will reject the null hypothesis that the means and/or distributions are the same. This point (and the previous point as well) can be illustrated by comparing the population birth weight distributions of the USA for children born in 1985 and 1986. Table C1 shows that there were over 3.7 million births in each year and the difference between the mean birth weights was 1.83 grams. Figure C1 shows the distributions graphically: 1985 is black and slightly thicker than 1986, which is light grey and thinner so that the two distributions are visible. There are some extremely minor differences between the two distributions, but both on means and on distributions the two populations are nearly identical.

[Table C1 and Figure C1 about here]

However, when we conduct inferential tests such as a t-test on the means and a Kolmogorov-Smirnov test to examine differences in the cumulative frequency distributions, the two populations are statistically different at very high levels of significance on both tests. The very large sample size has reduced the standard errors to such an extent that a mean difference of 0.003 standard deviations is statistically significant. From this, it is clear that statistical significance tests are not really appropriate for comparing the population data in the article. The differences in the means and distributions may be statistically significant, but the magnitudes of the difference are so small that this is meaningless for understanding the scientific significance (Ziliak and McCloskey, 2008). This is one of the potential problems with the ‘Big Data’ revolution. Thus, it is important to come to some *a priori* way of distinguishing what is a substantial, scientifically meaningful change in birth weight and what is not apart from looking at statistical significance (see below for more details).

The third reason that statistical tests are unhelpful for comparing birth weight distributions is that non-parametric tests that measure the equality of distributions (here simple Kolmogorov-Smirnov tests are used) are very sensitive to any differences in rounding or heaping that may occur between the groups being compared. This is especially important because the precision of measurement of birth weight has changed over time. In the United States in 1985, birth weight was measured to the nearest ounce. Drawing histograms with each bar being the width of an ounce, a quick look at the 1985 population of white singleton births in Boston (panel D in Figure C2) suggests that although there is a little bit of heaping, the distribution is smooth and precisely measured. However, the historical Boston birth weight samples in the other panels of Figure C2 were not measured with the same level of precision. The Lying-in Inpatient hospital was the most precisely measured since there were cases at every ounce increment. However, there is clear heaping on every half, quarter and even eighth of a pound because these values nearly always have more observations than their adjacent odd ounce measurements. The New England Hospital was measured with even less precision with heaping prominent on the half and quarter pound, though there were some children measured to the exact ounce. Finally, the Lying-in Outpatient Hospital seems to have only recorded birth weights to the nearest quarter pound but there is still significant heaping on the half pound intervals. The question then is how problematic are these rounding and heaping issues for comparing the distributions with a Kolmogorov-Smirnov Test?

[Figure C2 about here]

To test this, it is easiest to work with the comparator group to the historical samples, the population of white, singleton births in 1985. By simulating random rounding for a certain percentage of cases in this population and comparing the rounded population with the original, it is possible to see the extent of rounding necessary to reject the null hypothesis that the distributions are similar using a K-S test. There are 4,967 births in this population, so although the sample size is large, this exercise will be less strongly influenced by sample size than the example above on the entire population of US births.

To test how the threshold when rounding would influence the K-S test, an increasing percentage (ranging from 0.1% to 100%) of the Boston population was randomly rounded to the nearest quarter pound, half pound or full pound. Each new rounded distribution was then compared to the original in a K-S test to see when the differences became significant. Table C2 shows these results. Note that even as the percentage of the population rounded increases, the means of distributions do not change very much. If rounding occurred to the nearest quarter pound, 60% of the population must be rounded in order for the distribution to be significantly different from the original. For rounding to the nearest half pound, only 25% of the cases must be rounded for the distribution to be significantly different from the original. Finally, if rounding is occurring to the nearest pound, then only 15% of cases need to be rounded for the distribution to be statistically different. Figure C3 plots histograms for the critical rounding levels (panels A-C) and the original population.

[Table C2 and Figure C3 about here]

Because the New England Hospital and Lying-in Outpatient Ward are measured to the nearest quarter pound (and also are heaped), they will never appear similar to the modern, smooth data on a K-S test. The Lying-in Inpatient Ward has a better chance, but looking at the distributions in Figure C3, it is also very likely to fail a K-S test simply because of the heaping of measurements. Thus, the rounding and heaping in the historical data means that these distributions would always fail a K-S test against smooth, precise modern data even if the underlying, unrounded distributions were identical. One could try to adjust the 1985 data more carefully to match the heaping in the historical samples, but this involves a degree of 'massaging' the data that is more likely to raise questions than answer them. In addition, binning the 1985 data and comparing the samples using a Chi-square test would also be problematic because the historical data are not simply rounded. The degree of heaping would also lead to significant differences in the distributions. Thus, the fact that all of the historical distributions are statistically different from the 1985 population on a K-S test (Table C3) is not surprising or meaningful. We can learn more by looking for general patterns in the kernel density functions as were discussed in the paper.

[Table C3 about here]

As an aside, it is also important to note that the degree of precision in measurement also influences the low birth weight percentage of the sample. Because 2,500 grams, the low birth weight threshold, is equal to 5.51 pounds, any rounding or heaping on the half pound can lead to discrepancies in the low birth weight percentage. Table C2 shows that when one hundred percent of the 1985 population is rounded to the nearest quarter, half or full pound, then the low birth weight percentage varies by two percentage points from 5.16% to 7.15%. If heaping on the half pound were added to this, the distortion could be even greater.

Turning to the results of the statistical tests themselves, the means of all of the historical distributions are significantly different than the mean birth weight of white, singleton babies born in Boston in 1985 (based on one-sample t-tests presented in Table C3). However, two of the hospitals have mean birth weights that are 100 grams higher than the 1985 mean and one has a mean 50 grams below. How should this be interpreted? Have birth weights increased or decreased over time in Boston? All of the samples have mean birth weights above the median for the INTERGROWTH-21st standard, suggesting relatively good conditions. The weighted average of the three historical samples (3,440.57 grams) could be taken as a more representative mean, suggesting the mean birth weight had decreased by 63.79 grams between the two periods. Thus, by this measure birth weight decreased across a period where there were tremendous improvements in all other aspects of health.

However, this article has generally argued that birth weight distributions are similar in the past and present. This argument is based not on focussing on the statistical significance of the results but on the absolute magnitude of differences. The changes in mean birth weights that have occurred over the past 150 years are small relative to the difference in mean birth weights across countries. Even when the INTERGROWTH-21st study selected healthy women who had access to antenatal care and gave birth in hospitals, Indian children still had a mean term birth weight of

2.9 kg, 0.4 kg below their proscriptive standard. If the medical interventions and improvements in health over the past 150 years in the United States and Western and Northern Europe have not led to substantial increases in birth weight, will these Indian children converge to the standard? The changes in mean birth weight in the United States and Western and Northern Europe over time are also small relative to the secular increase in adult stature that took place over the same period. Mean adult stature has increased by 1.4 to 2.3 standard deviations whereas birth weight has only changed by an average of 0.18 standard deviations. If birth weight had kept up with increases in adult stature, the mean birth weight in the Netherlands should have increased by over 1,380 grams (Hatton and Bray, 2010).

In conclusion, this lengthy discussion of the statistical tests highlights the limitations of these methods for addressing how birth weights have changed over time. Rather than focussing on statistical significance, it is more helpful to analyse the difference in absolute magnitudes over time.

Web Appendix D: Data Sources

Bonellie, S., Chalmers, J., Gray, R., Greer, I., Jarvis, S., & Williams, C. (2008). Centile charts for birthweight for gestational age for Scottish singleton births. *BMC Pregnancy and Childbirth*, *8*(1), 512–10. http://doi.org/10.1186/1471-2393-8-5

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, University Lying-in Hospital, Montréal, 1843-1905. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/11272/MFJDL.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, Allgemeines Krankenhaus, Vienna, 1865-1930. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/11272/OSIEV.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, Utrecht Hospital, 1880-1940. Abacus Dataverse Network. Version 4 (2012), http://hdl.handle.net/11272/M0MUZ.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, Edinburgh Royal Maternity Hospital, 1847-1920. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/11272/YPAFV.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, The Rotunda (lying-in hospital), Dublin, 1869-1930. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/11272/JMBOK.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, New England Hospital for Women and Children, Boston, 1872-1900. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/2429/44366.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, University Lying-in Hospital, Montréal, 1843-1905. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/11272/MFJDL.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, Boston Lying-in (inpatient services), 1886-1900. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/2429/44366.

[dataset] Gagné, M. H., & Ward, W. P. Birth weight and economic growth data sets, Boston Lying-in (outpatient services), 1883-1900. Abacus Dataverse Network. Version 3 (2012), http://hdl.handle.net/2429/44366.

[dataset] Goldin, Claudia, and Robert A. Margo, 'Birthweight Data From the Philadelphia Almshouse Hospital, 1848-1873', ICPSR 20701, Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2008-01-03. http://doi.org/10.3886/ICPSR20701.v1.

[dataset] Haines, Michael R., 'Fetal death ratio, neonatal mortality rate, infant mortality rate, and maternal mortality rate, by race: 1850–1998', Table Ab912-927 in *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition,* edited by Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright (New York: Cambridge University Press, 2006). http://dx.doi.org/10.1017/ISBN-9780511132971.Ab912-113710.1017/ISBN-9780511132971.Ab912-1137.

Historisk Statistik för Sverige, Del 1. Befolkning, Andra Upplagan, 1720-1967 (Stockholm, 1967).

Macfarlane, Alison *et al.*, *Birth counts: Statistics of pregnancy and childbirth, Volume 2, tables* (London, 2000).

Martin Joyce A., Brady E. Hamilton, Michelle J. K. Osterman *et al.*, 'Births: Final data for 2012', *National vital statistics reports*, 62, no 9 (2013), available at http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_09.pdf.

[dataset] National Center for Health Statistics (1986). Data File Documentations, Natality, 1986 (machine readable data file and documentation), National Center for Health Statistics, Hyattsville, Maryland. Available at http://www.nber.org/data/vital-statistics-natality-data.html.

[dataset] National Center for Health Statistics (2013). Data File Documentations, Natality, 2013 (machine readable data file and documentation), National Center for Health Statistics, Hyattsville, Maryland. Available at http://www.nber.org/data/vital-statistics-natality-data.html.

National Statistics for Scotland. (2014). Births in Scottish Hospitals. Table 5. http://www.isdscotland.org/Health-Topics/Maternity-and-Births/Publications/2014-08-26/mat_bb_table5.xls

[dataset] Office of National Statistics, 'Birth Summary Tables, England and Wales 2013', (2014), electronic dataset, accessed 4/6/2015, http://www.ons.gov.uk/ons/rel/vsob1/birth-summary-tables--england-and-wales/2013/index.html.

Registry Department, City of Boston, *Annual Report of the Registry Department of the City of Boston for the Year 1907* (Boston, 1908).

[dataset] Statistics Norway, 'Table: 05860 – Perinatal deaths and infant deaths', StatBank Norway, accessed 14/10/2014, https://www.ssb.no/en/befolkning.

U.S. Department of Commerce, Bureau of the Census, *Birth, Stillbirth, and Infant Mortality Statistics for the Birth Registration Area of the United States, 1931: Seventeenth Annual Report* (Washington, 1934).

[dataset] U.S. Department of Health and Human Services, National Center for Health Statistics, 'Linked birth/infant death data, 1985 birth cohort: United States', ICPSR 3266. Hyattsville, MD: U.S. Dept. of Health and Human Services, National Center for Health Statistics [producer], 1990. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2003. http://doi.org/10.3886/ICPSR03266.v1.

U.S. Department of Health, Education and Welfare, Public Health Service, *Vital Statistics of the United States, 1950: Volume II, Marriage, Divorce, Natality, Fetal Mortality and Infant Mortality Data* (Washington, 1953).

U.S. Department of Health, Education and Welfare, Public Health Service, *Vital Statistics of the United States, 1960: Volume 1 – Natality* (Washington, 1960).

U.S. Department of Health, Education and Welfare, Public Health Service, *Vital Statistics of the United States, 1970: Volume 1 – Natality* (Rockville, Maryland, 1975).

U.S. Department of Health, Education and Welfare, Public Health Service, *Vital Statistics of the United States, 1980: Volume 1 – Natality* (Hyattsville, Maryland, 1984).

U.S. Department of Health, Education and Welfare, Public Health Service, *Vital Statistics of the United States, 1990: Volume 1 – Natality* (Hyattsville, Maryland, 1994).

U.S. Department of Health, Education and Welfare, Public Health Service, *Vital Statistics of the United States, 2000: Volume 1 – Natality*, online publication, http://www.cdc.gov/nchs/data/statab/t001x26.pdf.

[dataset] Ward, W. P., Jahanfar, S. & Lesack, P. Bologna Sant'Orsola Maternity Clinic Database, 1880-1940. Version 6 (2014), http://hdl.handle.net/11272/10050.

New References in Web Appendix

Ward, W. P. (2004). Perinatal mortality in Bologna, 1880-1940. In Breschi, M., & Pozzi, L. (Eds.). *The Determinants of Infant and Child Mortality in Past European Populations* (pp. 213-230). Udine: Forum.

Ward, W. P. (2003). Perinatal mortality in Utrecht, The Netherlands, 1880–1940. *Economics and Human Biology*, *1*(3), 379–398. http://doi.org/10.1016/S1570-677X(03)00070-4

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

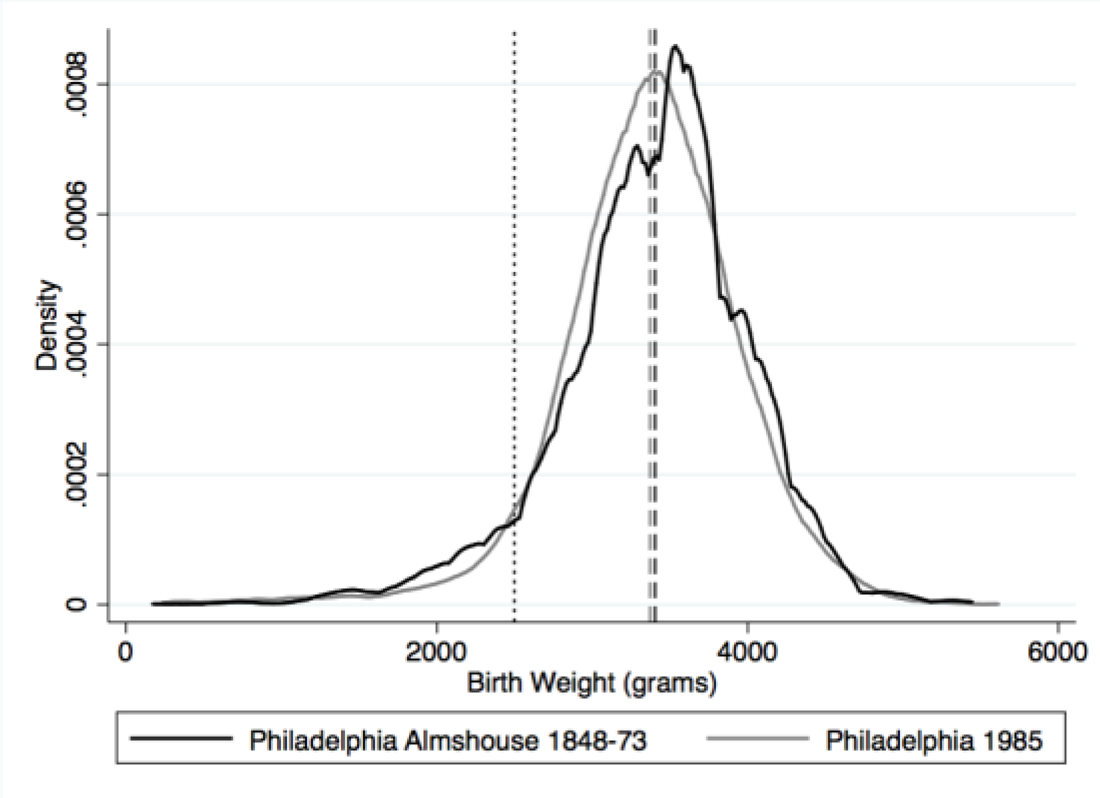
Web Appendixes – Figures and Tables

Table A1: Birth weights of children around the world in the 1980s.

Country	Mean Birth Weight (g)	LBW (%)
Latin America		
Brazil	3,170-3,298	9.0
Chile	3,340	9.0
Colombia	2,912-3,115	10.0
Guatemala	3,050	17.9
Mexico	3,019-3,025	11.7
Africa		
Egypt	3,200-3,285	7.0
Kenya	3,143	12.8
Nigeria	2,880-3,117	18.0
Tunisia	3,210-3,376	7.3
United Republic of Tanzania	2,900-3,151	14.4
Zaire	3,163	15.9
Asia		
China	3,215-3,285	6.0
India	2,493-2,970	30.0
Indonesia	2,760-3,027	14.0
Iran	3,012-3,250	14.0
Iraq	3,540	6.1
Japan	3,200-3,208	5.2
Malaysia	3,027-3,065	10.6
Pakistan	2,770	27.0

Sources: Kramer (1987, p. 665).

Figure A1: Birth weights in the Philadelphia Almshouse (1848-73) compared to the population distribution of births in Philadelphia in 1985.



Notes: The dotted vertical line marks the low birth weight cut-off of 2,500 grams. The dashed vertical lines mark the mean of each distribution. The distribution of birth weights in Phildadelphia in 1985 only includes white, singleton births to make it most comparable with the historical data.

Sources: Philadelphia Almshouse – Goldin and Margo (2008); Birth weights in Boston 1985 – U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990).

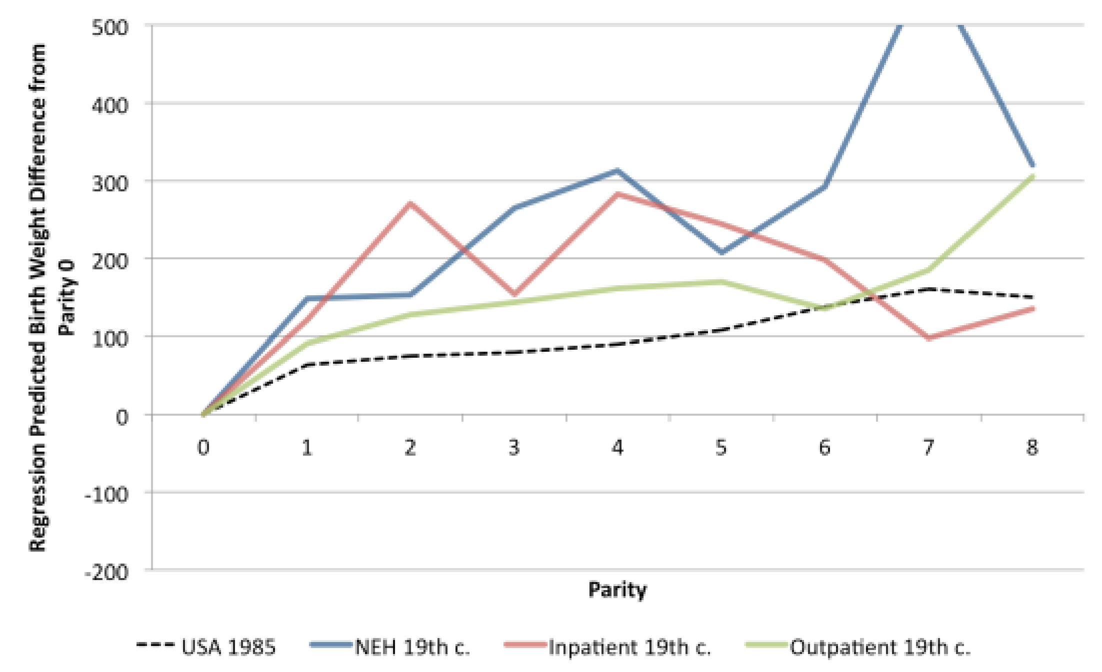
Table A2: Estimates of the effect of tobacco smoke exposure on average birth weight in the United States.

	Smoking Exposure				Predicted Birth Weights	
	None	Passive	Low	High	Mean	Change
Birth Weights (Roquer et al., 1995)	3,407	3,215	3,205	2,948		
Share of Mothers Exposed						
Minimum Smoking (c. 1900)	90%	10%	0%	0%	3388	
Maximum Smoking (c. 1970)	30%	30%	30%	10%	3243	-145
2013 USA	82%	10%	5%	4%	3361	118

Notes: Predicted mean birth weights are the weighted average of Roquer *et al.*'s figures based on the assumed shares in each category.

Sources: Roquer *et al.* (1995); National Center for Health Statistics (2013).

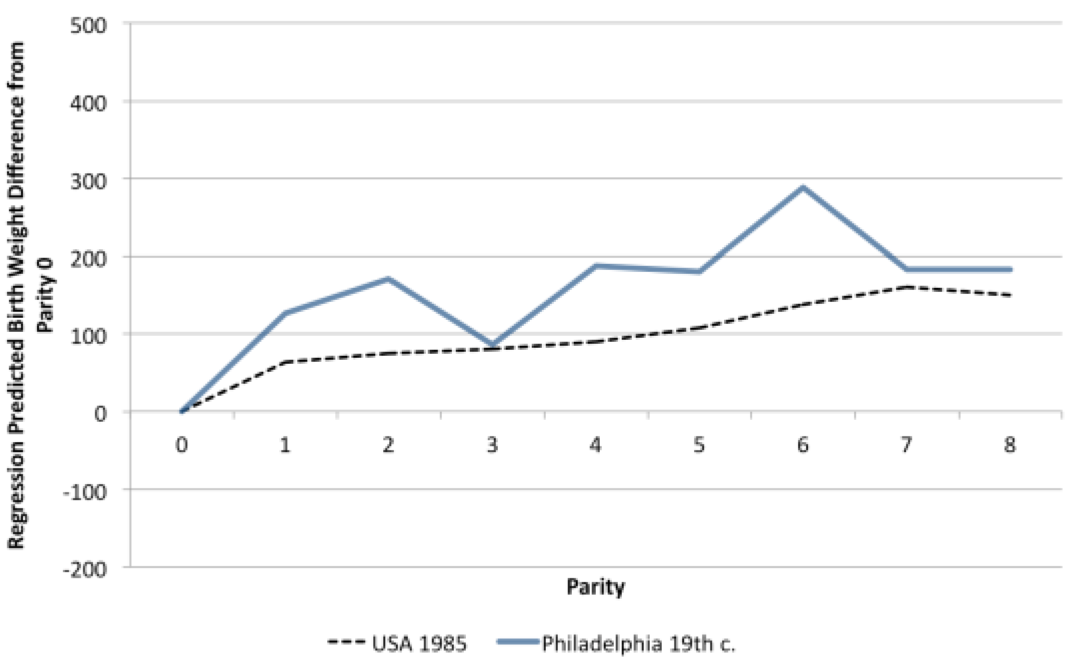
Figure A2: Regression predicted birth weight difference from parity zero for hospitals in nineteenth-century Boston compared to the US figure in 1985.



Notes: The birth weight difference at each parity was estimated using multiple regression analysis with a dummy variable for children of each parity. The regressions also included dummy variable controls for maternal age, sex of the child and gestational age where possible.

Sources: Historical birth weight datasets – Gagné and Ward (2012); Birth weights in Boston 1985 – U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990).

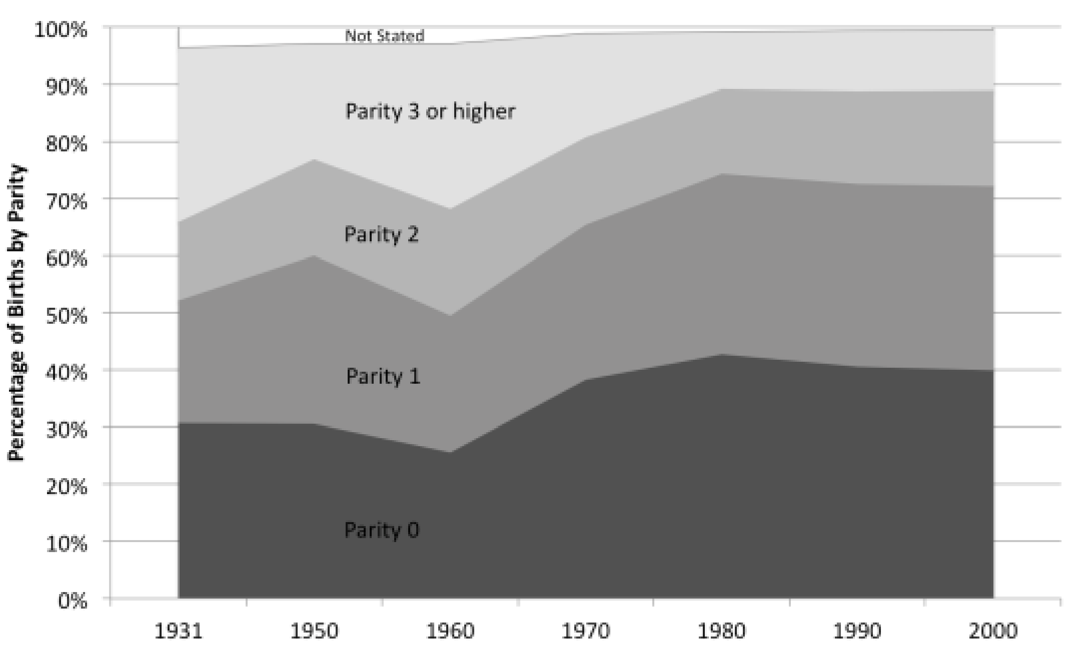
Figure A3: Regression predicted birth weight difference from parity zero for the Philadelphia Almshouse (1848-73) compared to the US figure in 1985.



Notes: The birth weight difference at each parity was estimated using multiple regression analysis with a dummy variable for children of each parity. The regressions also included dummy variable controls for maternal age, sex of the child and gestational age where possible.

Sources: Historical birth weight datasets – Philadelphia Almshouse – Goldin and Margo (2008); Birth weights in Boston 1985 – U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990).

Figure A4: Change in share of births by parity in the United States, 1931-2000.



Notes: Data are from the entire birth registration area across the dates covered.

Sources: Vital Statistics of the United States 1931-2000.

Table B1: Important statistics for the available maternity hospital samples in the nineteenth century.

City	Hospital	Years	Mean Birth Weight (g)	% Low Birth Weight	Sample Size
North America					
Boston	New England Hospital	1872-1900	3,480	5.79%	3,109
Boston	Lying-in: Inpatient Ward	1886-1900	3,330	6.55%	2,261
Boston	Lying-in: Outpatient Ward	1884-1900	3,479	4.34%	3,294
Philadelphia	Philadelphia Almshouse	1848-73	3,406	7.46%	4,448
Montreal	University Lying-in Hospital	1847-99	3,403	6.48%	5,155
Europe					
Bologna	Ospedale Sant'Orsola	1880-99	3,011	16.16%	2,772
Dublin	The Rotunda	1869-99	3,203	10.82%	4,816
Edinburgh	Edinburgh Royal Maternity	1847-99	3,227	11.05%	6,216
Norway	Oslo, Bergen and Trondheim	1860-1900	3,400-3,500	N/A	N/A
Utrecht	Algemeen Ziekenhuis	1880-99	3,252	6.14%	2,051
Vienna	Allgemeines Krankenhaus	1865-99	3,098	9.32%	5,825

Notes: Stillbirths have been excluded from all samples. N/A means that the information was not available in the data file or published study. Shaded hospitals have been excluded from the analysis because there is ample evidence of selection on poor health, not just on socioeconomic background.

Sources: see table B2.

Table B2: Information on the characteristics of the various historical maternity hospitals

City	Hospital	Years	% of Population Births in Hospital	Class Served	% Primi-parous Mothers	% Unmarried	Fee?	Binding Fee	Moral Admissions Criteria	Selection on Poor Health	Sources
North America											
Boston	New England Hospital	1872-1900	c. 2.5%	Low mid, upper working	55.6%	30.5%	Yes	Yes	Yes	No	c, g
Boston	Lying-in: Inpatient Ward	1886-1900	c. 2.5%	Poorer working	64.6%	42.9%	Yes	No	Yes	No	c, h
Boston	Lying-in: Outpatient Ward	1884-1900	c. 10 %	Working	19.1%	2.2%	No	No	No	No	c, i
Philadelphia	Philadelphia Almshouse	1848-73	N/A	Poor working	60.0%	56.3%	No	No	N/A	No	a, j
Montreal	University Lying-in Hospital	1847-99	c. 2-3%	English-speaking working	60.1%	65.0%	No	No	No	No	c, k
Europe											
Bologna	Ospedale Sant'Orsola	1880-99	c. 5-10%	Working	31.2%	22.3%	No	No	No	Yes	f, l
Dublin	The Rotunda	1869-99	c. 15%	Working class and poor	35.9%	4.3%	No	No	No	No	c, m
Edinburgh	Edinburgh Royal Maternity	1847-99	c. 3-5%	Poor working	40.7%	73.5%	No	No	No	No	c, n
Norway	Oslo, Bergen and Trondheim	1860-1900	Low	Working	N/A	c. 67%	N/A	N/A	N/A	No	b
Utrecht	Algemeen Ziekenhuis	1880-99	c. 4%	Working class bias	61.0%	c. 75%	Yes	No	No	Yes	d, o
Vienna	Allgemeines Krankenhaus	1865-99	c. 20%	Poor and working	48.4%	95.0%	No	No	No	No	c, p

Notes: N/A means that the information was not available in the data file or published study. Shaded hospitals have been excluded from the analysis because there is ample evidence of selection on poor health, not just on socioeconomic background. Percentages of primiparous women and unmarried mothers may differ from the published articles because these statistics were calculated for a different period and only for the mothers of children whose birth weight was recorded.

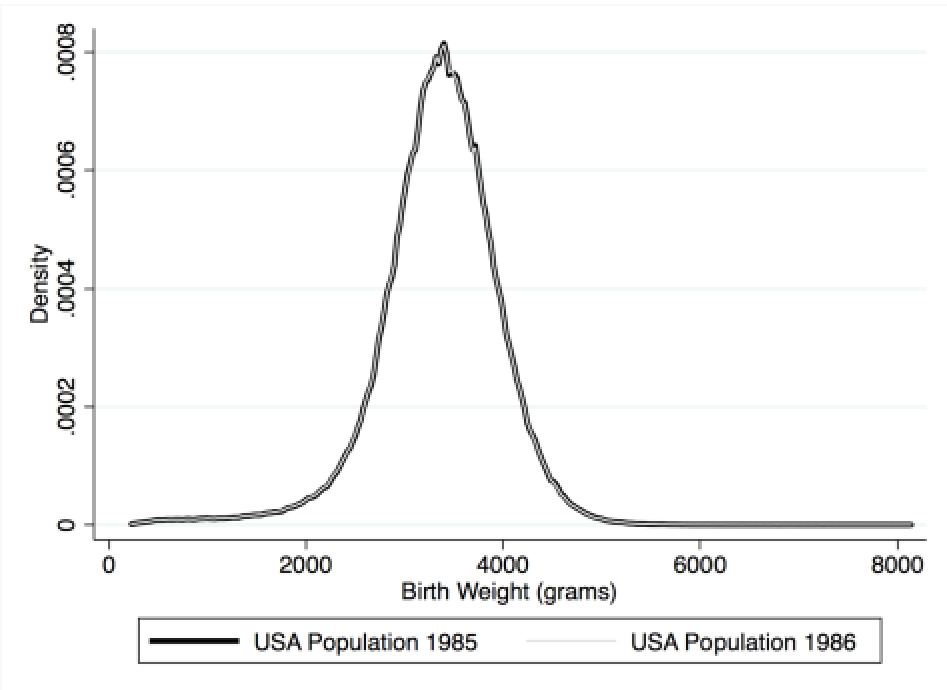
Sources: Published sources: a – Goldin and Margo (1989); b – Rosenberg (1988); c – Ward (1993); d – Ward (2003); f – Ward (2004). Data sources: g-i and l-p – Gagné and Ward (2012); j – Goldin and Margo (2008).

Table C1: Comparison of US population of birth weights in 1985 and 1986.

Data	N	Mean	Standard Deviation	Standard Error
USA 1985	3,760,543	3,350.32	601.98	0.31043
USA 1986	3,755,413	3,348.49	602.07	0.31068
Mean Difference		-1.83		
Mean Difference as % of SD		-0.304%		
Statistical Tests				
T-Test on Mean Difference				
T-value		4.1742		
P-value		<0.0000		
Kolmogorov-Smirnov Distribution Test				
P-value		0.001		

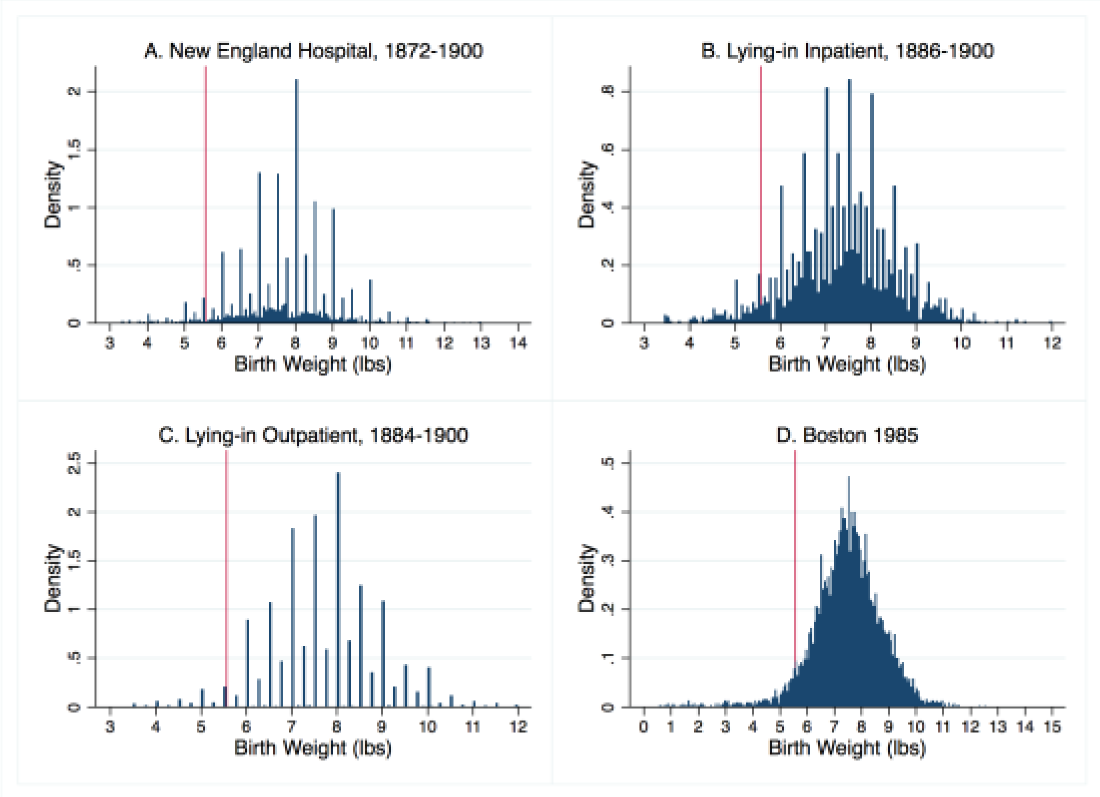
Sources: U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990); National Center for Health Statistics (1986).

Figure C1: Comparison of US population birth weight distributions, 1985 and 1986



Sources: see Table C1.

Figure C2: Histograms showing the density of observations at one ounce intervals across the range of birth weights for the historical sample and modern population.



Notes: The red vertical line marks the low birth weight cut-off of 2,500 grams. The distribution of birth weights in Boston in 1985 only includes white, singleton births to make it most comparable with the historical data.

Sources: Historical birth weight datasets – Gagné and Ward (2012); Birth weights in Boston 1985 – U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990).

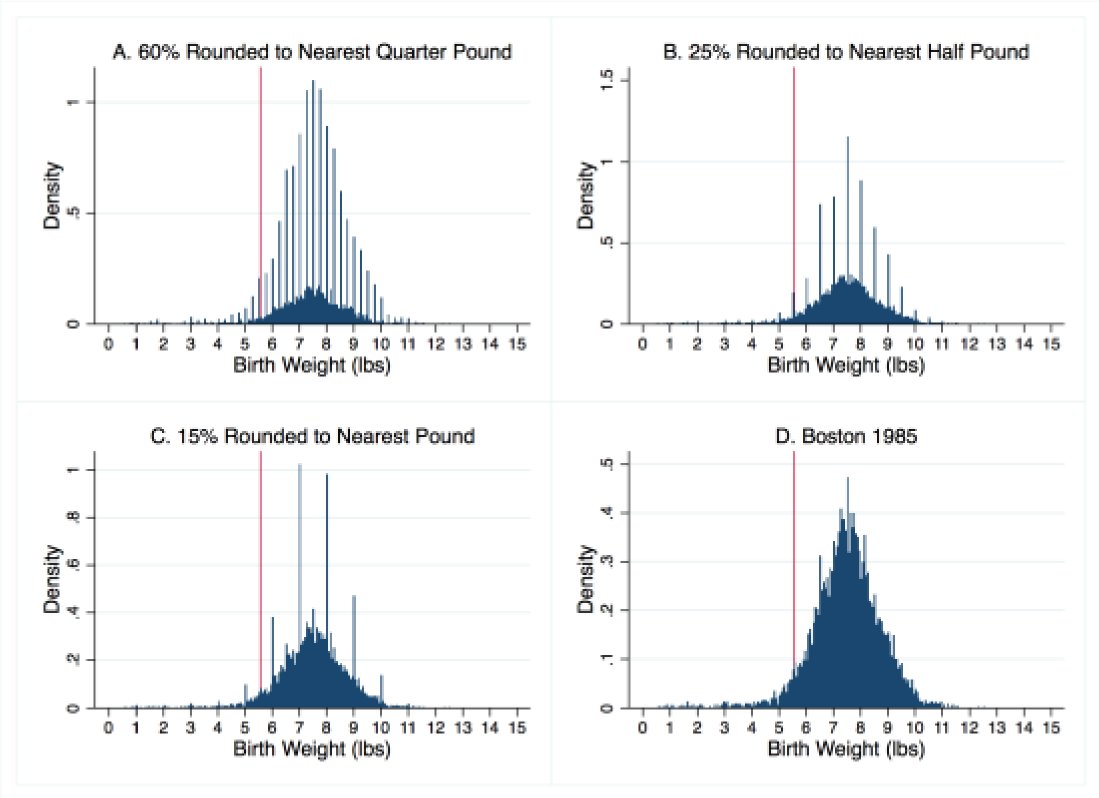
Table C2: Comparisons of various randomly rounded versions of the 1985 Boston population of births with the unrounded version.

Percentage of Pop. Rounded	KS Test P-values			Mean Birth Weight (ounces)			Low Birth Weight %		
	1/4 lb	1/2 lb	1 lb	1/4 lb	1/2 lb	1 lb	1/4 lb	1/2 lb	1 lb
0.1%	1.000	1.000	1.000	119.11	119.11	119.11	5.64%	5.64%	5.64%
1%	1.000	1.000	1.000	119.11	119.12	119.12	5.64%	5.64%	5.64%
10%	1.000	0.940	0.312	119.16	119.15	119.18	5.66%	5.78%	5.56%
15%	1.000	0.657	0.043	119.19	119.19	119.23	5.74%	5.86%	5.58%
20%	0.921	0.300	0.002	119.22	119.21	119.22	5.76%	5.92%	5.56%
25%	0.848	0.041	0.000	119.23	119.26	119.28	5.84%	5.94%	5.52%
30%	0.657	0.031	0.000	119.25	119.27	119.29	5.76%	5.97%	5.48%
40%	0.300	0.002	0.000	119.31	119.30	119.27	5.97%	6.23%	5.46%
50%	0.095	0.000	0.000	119.36	119.34	119.40	5.97%	6.63%	5.40%
60%	0.013	0.000	0.000	119.41	119.41	119.48	6.03%	6.65%	5.30%
75%	0.003	0.000	0.000	119.48	119.48	119.55	6.13%	6.81%	5.30%
100%	0.000	0.000	0.000	119.61	119.61	119.70	6.23%	7.15%	5.16%

Notes: The distribution of birth weights in Boston in 1985 only includes white, singleton births to make it most comparable with the historical data. Shaded values are statistically significant at the 5% level.

Sources: U.S. Department of Health and Human Services, 'Linked birth/infant death data, 1985 birth cohort' (1990).

Figure C3: Histograms showing the density of observations at one ounce intervals across the range of birth weights for the randomly rounded and original 1985 population.



Notes: The red vertical line marks the low birth weight cut-off of 2,500 grams. The distribution of birth weights in Boston in 1985 only includes white, singleton births to make it most comparable with the historical data.

Sources: see Table C2.

Table C3: Descriptive statistics and inferential tests of means and distributions of historical birth weight samples and the 1985 population.

	Boston 1985	New England Hospital	Lying-in Inpatients	Lying-in Outpatients
Descriptive Statistics				
Mean	3,376.78	3,480.16	3,329.73	3,479.29
Mean Difference from Boston 1985		103.38	-47.05	102.51
SD	586.51	581.03	524.19	558.25
N	4,967	3,109	2,261	3,294
One-sample T-test on 1985 Mean				
T-value		9.92	-4.27	10.54
P-value		0.00	0.00	0.00
Kolmogorov-Smirnov Test				
P-value		0.00	0.00	0.00

Sources: see Figure C2.