Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Obstetrics & Gynecology and Reproductive Biology 111 (2003) S33–S44
EUROPEAN JOURNAL OF OBSTETRICS & GYNECOLOGY AND REPRODUCTIVE BIOLOGY
www.elsevier.com/locate/ejogrb

Methodological difficulties in the comparison of indicators of perinatal health across Europe

Nicholas Lack[a], Jennifer Zeitlin[b,*], Lone Krebs[c], Wolfgang Künzel[d], Sophie Alexander[e]

[a]*Bereichsleitung Peri-/Neonatologie, Bayerische Arbeitsgemeinschaft Qualitätssicherung in der stationären Versorgung (BAQ)—Bavarian Quality Assurance for In-Patient Medical Care, Munich, Germany*
[b]*Epidemiological Research Unit on Perinatal and Women's Health, INSERM U149, Paris, France*
[c]*Department of Obstetrics, University of Copenhagen, Copenhagen, Denmark*
[d]*Department of Obstetrics and Gynaecology, Frauenklinik der Justus Liebig, University of Giessen, Hesse, Germany*
[e]*Reproductive Health Unit, School of Public Health, Université Libre de Bruxelles, Brussels, Belgium*

Abstract

The main purpose of this article is to point out common pitfalls that can confuse comparative analyses of indicators of perinatal health and to discuss ways to overcome or minimize these difficulties. The challenge is to distinguish 'real' variations in the value of an indicator from variations due to differences in registration practices and definitions and from random variation. The first section presents the major properties that are desirable in indicators of perinatal health status and perinatal health care in Europe to be used for comparative purposes. The second section provides specific examples of the types of methodological difficulties encountered in European cross-country comparisons due to variations in the definition, measurement and construction of indicators. The conclusion discusses the PERISTAT project's responses to these difficulties and how these methodological constraints impact on the selection of an appropriate indicator set for Europe today.

Keywords: Perinatal health; Indicators; Perinatal mortality; Registration practises; Data collections systems; Europe

Confronted with the data in Table 1 showing the distribution of fetal mortality rates in selected European countries, the critical reader is automatically torn between the urge to interpret the differences in the light of pre-existing ideas about the determinants of perinatal health in the member states, on the one hand, and a more cautious and detached critical stance, on the other, with the knowledge of a multitude of confounding aspects to which the apparent differences in observed rates may be attributed. Differences in health outcomes may reflect the varying demographic characteristics of women of childbearing age in different European countries, or differences in social conditions, or the considerable spread in per capita expenditures on health care, e.g. Germany US$ 2713, Finland US$ 1789, Spain US$ 1071 [1]. Conversely, the inter-country differences in registration systems, for example also imply biases in recorded mortality rates. The very essence of Europe lies in its heterogeneous composition. For perfect comparability, the recording systems for measuring perinatal health should ideally be identical. Although perfect comparability is inherently unattainable, it may at least be approximated.

The main purpose of the present article is to point out common pitfalls that can confuse comparative analyses of indicators and to discuss ways to overcome or minimize these difficulties. The challenge is to distinguish 'real' variations in the value of an indicator from variations due to differences in registration practices and definitions and from random variation. The type of data sources used to construct indicators also affects their comparability, but this issue is discussed in Macfarlane et al. (this issue) and will not be addressed below.

This paper is divided into two principal sections. We start by discussing the major properties that are desirable in indicators of perinatal health status and perinatal health care in Europe for comparative purposes. The second section provides examples of the types of methodological difficulties encountered in European cross-country comparisons due to variations in the definition, measurement and construction of indicators. In our conclusion, we discuss the PERISTAT project's responses to these difficulties and how these methodological constraints affected the proposed

* Corresponding author. Tel.: +33-142345579.
E-mail address: zeitlin@cochin.inserm.fr (J. Zeitlin).

0301-2115/$ – see front matter © 2003 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ejogrb.2003.09.004

Table 1
Fetal death rate, data provided to the PERISTAT project

Member state (coverage)	Source	Fetal death rate*
Austria	A1-2001	3.6
Belgium (Flanders)	B2-2000	4.5
Denmark	DK1-2000	3.8
Finland	FIN1-2000	4.0
France	F2-2000	4.6
Germany (nine Bundesländer)	D1-2000	3.7
Greek (perinatal survey)	EL1-1998	5.7
Ireland	IR1-1999	5.3
Italy	I-1998	4.4
Luxembourg	L1&2-2000	5.5
The Netherlands	NL1-1999	7.4
Portugal	P1-1999	5.8
Spain (Madrid, Valencia, Pais Vasco)	E2-2000	5.2
Sweden	S1-2000	3.9
UK: England & Wales	UK1,5,17-2000	5.3
UK: northern Ireland	UK7-2000	4.4
UK: Scotland	UK6-2000	5.7

Note: data sources in Appendix A, described in Macfarlane et al. (this issue).

* Rates per 1000 live births and fetal deaths.

indicator set. Throughout, we shall refer to the core and recommended indicator sets proposed by the PERISTAT scientific advisory committee.

1. Optimal properties of indicators of perinatal health and care

It is good practise to begin by determining the characteristics of ideal indicators and then to use these as guidelines for selecting or constructing performance measures. Generally, good indicators are expected to be representative, useful, precise, highly sensitive, readily accessible, valid, specific, unbiased and reliable. Last but not least, the entire set of indicators should be comprehensive, that is, it should cover all aspects of the field of interest.

Indicators of health and health care may be viewed in terms of: (a) what they stand for, in terms of their content, (b) their formal characteristics and (c) practical concerns, such as cost and accessibility. The first of these three approaches refers to their general usefulness, their importance and comprehensiveness in assessing perinatal health and care. The second approach looks at technical qualities such as representativeness, validity and sensitivity. The third refers to practical concerns—the cost and effort involved in actually obtaining relevant and comparable data from various sources in different countries.

1.1. Criteria related to content

The indicator set proposed in this special issue has emerged from a Delphi process and formally reflects the needs seen by the scientific advisory committee, a representative body with members from all participating member states. The areas it covers are thus clearly considered useful and important, but probably constrained by their awareness of what is available. The division of this set into core indicators, recommended indicators and indicators proposed for future research further implies a ranking according to importance.

If the principal use of a national level indicator is as a 'marker of progress' towards improved reproductive health status, the most *useful* indicator would be one that is either a direct or proxy measure of the impact of healthcare services and policies. Many of the suggested impact indicators measure mortality, which is only a very partial measure of health outcome. These indicators may miss important variations in reproductive morbidity and may be of limited use as measures of change in overall reproductive health status. *Comprehensive coverage* by a set of indicators is essential for a complete assessment and comparison of perinatal health across different countries. Focusing merely on readily available data on mortality rates and birth weight distributions may yield reliable information but will not sufficiently describe the state of perinatal health in any of the countries considered. An adequate description of perinatal health will require a *balanced indicator set* that preferably covers mutually exclusive domains.

1.2. Criteria related to formal characteristics

The formal requirements for indicators are easiest to define, and they therefore exist in abundance. Popular choices are described hereafter. These prerequisites are ideals and are rarely achieved in practice, however. Very few if any indicators will ever match up 100% to all these properties. In that sense they are rather to be seen as a checklist for choosing and assessing indicators of perinatal health.

A *valid* indicator is one that actually measures the issue or factor it is supposed to measure. Therefore, an essential starting point is to establish exactly what the indicator is supposed to be measuring. A *sensitive* indicator is one that can reveal changes in the issue or factor of interest. Indicators concentrating on mortality rates are rather insensitive to changes in overall reproductive health status since a substantial shift may occur in the burden of perinatal morbidity before it is reflected in mortality rates. When relatively small numbers are involved, there will be wide random variations in values and wide confidence intervals. While measures of more common events (such as different types of maternal morbidity) would be more sensitive to change, they still present measurement challenges. In this sense, precision is a necessary prerequisite for sensitivity. A *precise* indicator is stable: its variance is low. Typically, relative frequencies computed as rates of such rare events as maternal deaths have high variances that lead to inherent variations across countries and over time.

A *specific* indicator is one that reflects *only* changes in the issue or factor under consideration. Observed differences in

the values of an indicator may not reflect true differences in health status but may be influenced by a number of other artefactual or confounding factors. For example, observed changes may be due to improvements in the ascertainment of events such as maternal or perinatal deaths through the development of better reporting systems over time or may be due to differences in the case-mix characteristics of the population under study. This includes differences in the population age/sex case mix for crude birth rates or in the severity of cases for facility-based case fatality rates.

An *unbiased* indicator measures the quality of interest without any systematic distortions. Differing standards of legislation for the recording of fetal deaths may lead to biases in the computed rates, depending on the stipulated minimum lower limit adopted for recording. A *representative* indicator is one that adequately encompasses all the issues or population groups it is expected to cover. At the national level the group of interest is the entire population, including minority groups and adolescents. The representativeness of a given indicator will be compromised if there is selection bias either in the denominator defined for the indicator or in the source of the data used to generate it. A *reliable* indicator is one that will yield the same value if it is measured in the same way on the same population at almost the same time. Birth weight is generally recorded reliably, while individual satisfaction with the care received is affected by considerable recall bias if surveyed repeatedly.

1.3. Criteria related to practical concerns

Practical concerns bear on accessibility and cost. An *accessible* indicator is one for which the data required are already available or relatively easy to acquire by feasible survey methods already validated in field trials. Sources of information include vital registration, routine health service data, health service surveys, population-based surveys and surveillance. Indicators generated by routinely collected data are usually the most readily accessible but there may be serious problems with the representativeness and reliability of the data. To correct for these drawbacks, additional adjustments may be required; these may include remapping from ICD9 to ICD10 classifications, recoding socio-economic status to achieve comparability across Europe or changes in computations, for example for gestational age after further ultrasonic corrections. All this increases the effort and costs involved in providing comparable indicators.

Expecting a single indicator or even an entire set of indicators to embody all the above-mentioned qualities of content, formal characteristics and practical concerns can rapidly lead to the unsatisfactory situation where discrepancies between the ideal indicator set and the available information prevent the acceptance of any suitable indicators. From a practical point of view, a compromise must be struck between useful, important indicators that satisfy many of the formal characteristics and are still accessible. The task then entails taking the proposed indicators as given and dealing with the methodological problems that occur when they are applied in comparisons. This is the approach adopted in the following section, where we take the indicator set (core, recommended and future use) proposed by the scientific advisory committee as given. Hence, we will refer frequently to the tables compiled by the PERISTAT steering group.

2. Variation in Europe-wide indicators due to differences in registration, definition, measurement and precision: examples

European indicators do not always fulfil all the desired criteria described in Section 1: differences in health systems and health reporting systems throughout Europe present major challenges to the construction of comparable indicators. The following examples illustrate the impact of differences in the measurement and construction of these indicators on their values.

2.1. Differences in registration of perinatal deaths

Mortality indicators are particularly sensitive to biases related to the construction of indicators. For example, changes in birth notification and registration practices can cause major biases. In 1994 Germany reduced the lower limit for birth weight for registration of fetal deaths from 1000 to 500 g. As a consequence, the perinatal mortality rate jumped suddenly from 5.5 per 1000 to 6.6 per 1000, an increase of 20% (Fig. 1): mortality increases as birth weight decreases and comparatively small changes at the lower limit of birth weight lead to large changes in mortality rates. Macfarlane et al. in this special issue provide information on the lower limits for registration of stillbirths and live births in the EU countries.

The EURONATAL study compared perinatal mortality rates from 15 European countries and analysed these differences [2]. It demonstrated that the substantial differences in published perinatal mortality rates between western European countries are due to a large extent to the use of different cut-off points for birth weight and gestational age. Their analysis applied two methods of standardisation (direct and indirect) and used information on birth weight, gestational age and perinatal death from the Finnish database of 190,000 births over 3 years. In both cases the standardisation reduced the variation between countries and also changed their perinatal mortality rate rank order slightly.

Differences in registration that lead to more complete registration of the neonatal deaths of extremely preterm babies can also explain shifts in the timing of mortality, since these babies tend to die immediately after birth. Fig. 2 displays ratios of early to late neonatal deaths plotted over time to illustrate this impact. Overall we observe a trend that shifts mortality from the early neonatal period (within the

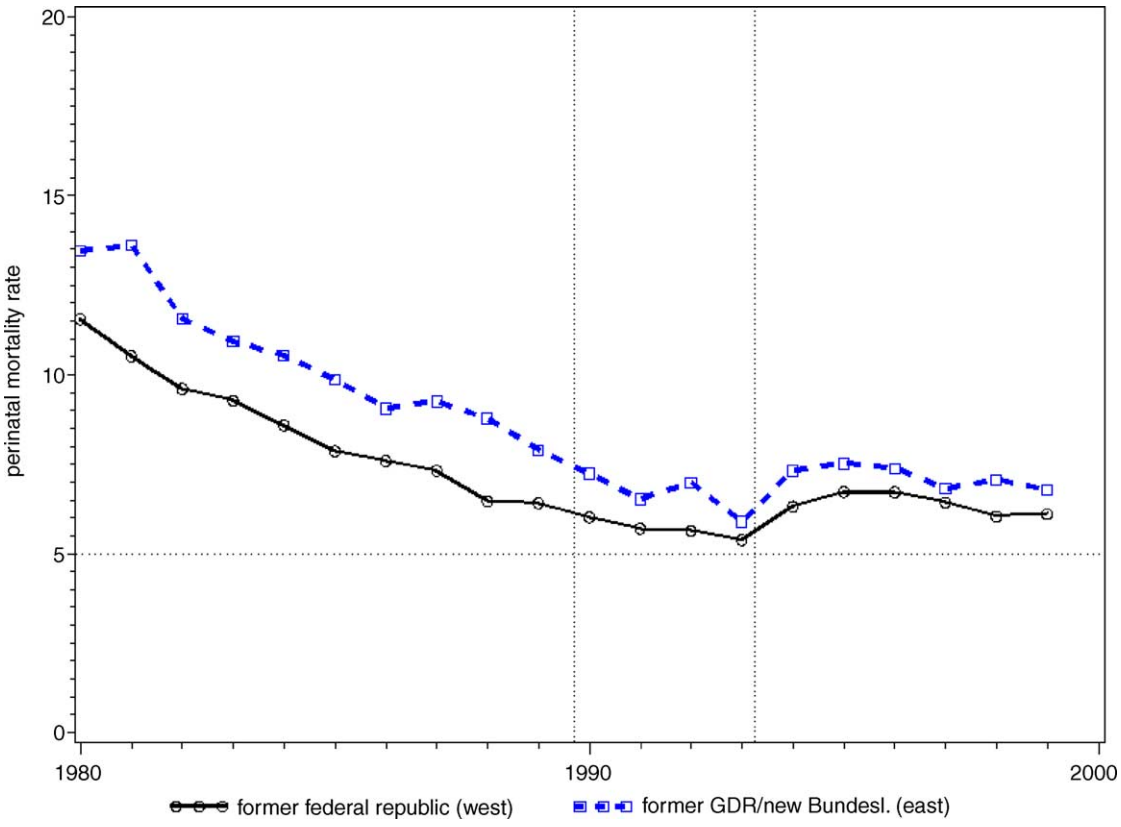

Fig. 1. Perinatal mortality in Germany (births in west and east Germany). German Reunification in October 1989. Minimum birth weight for official registration purposes was lowered for fetal deaths to 500 g from 1000 g with effect on 1 April 1994.

first 7 days of life) to the late neonatal period (8–28 days). These time series are not, however, free of trend reversals or sudden jumps. In Germany, for example, the effects of the reunification process are quite marked. The time series for Germany is bipartite, with only the federal republic (West Germany) plotted up until 1989, and the reunified state plotted from 1991 onwards. Similarly, legislation governing the registration of perinatal deaths changed in Italy in 1978, in France in 1975 and in 1994; these changes affected the registration of births at the limit of viability and thus increased the ratio of early to late neonatal deaths. Arrows on the graph indicate these increases. Another item that needs to be clearly defined is whether late pregnancy terminations for major fetal anomalies or for life-threatening maternal conditions are included in fetal deaths or omitted.

2.2. Differences in definition and measurement of gestational age

Preterm birth is one of the principal complications of pregnancy and has a significant impact on mortality and morbidity. This indicator is appealing for cross-country comparisons because preterm birth has a commonly agreed-upon definition (preterm babies are those born before 37 completed weeks of gestation), unlike growth restriction, which is measured against a variety of standards.

The exact duration of gestation, however, is rarely known. Before the development of ultrasound, the measurement of gestational age relied primarily on the pregnant woman's recall of the first day of her last menstrual period (LMP) or other clinical data (records of temperature, clinical assessment of maturity at birth). Recall error and the effect of irregular menstrual cycles made these measures quite imprecise. A significant proportion of the data was often missing or clearly erroneous.

The introduction of earlier prenatal visits improved the reliability of women's recollection of their LMP. More significantly, however, the spread of ultrasound has made it possible to obtain reliable estimates of gestational age when menstrual dates are uncertain and when cycles are irregular. Ultrasound can also be used to ''correct'' gestational age even when the LMP is known, when the ultrasound measures are not concordant with the gestational age derived from LMP calculations.

The use of ultrasound to estimate gestational age has an important impact on the distribution of gestational age in the population. The use of ultrasound decreases the proportion of reported postterm pregnancies (pregnancies ending at 42 weeks or later) and increases the proportion of preterm babies [3]. This is explained by a systematic shift associated with the greater frequency of late, compared with early, ovulation. Calculation of gestational duration based on LMP

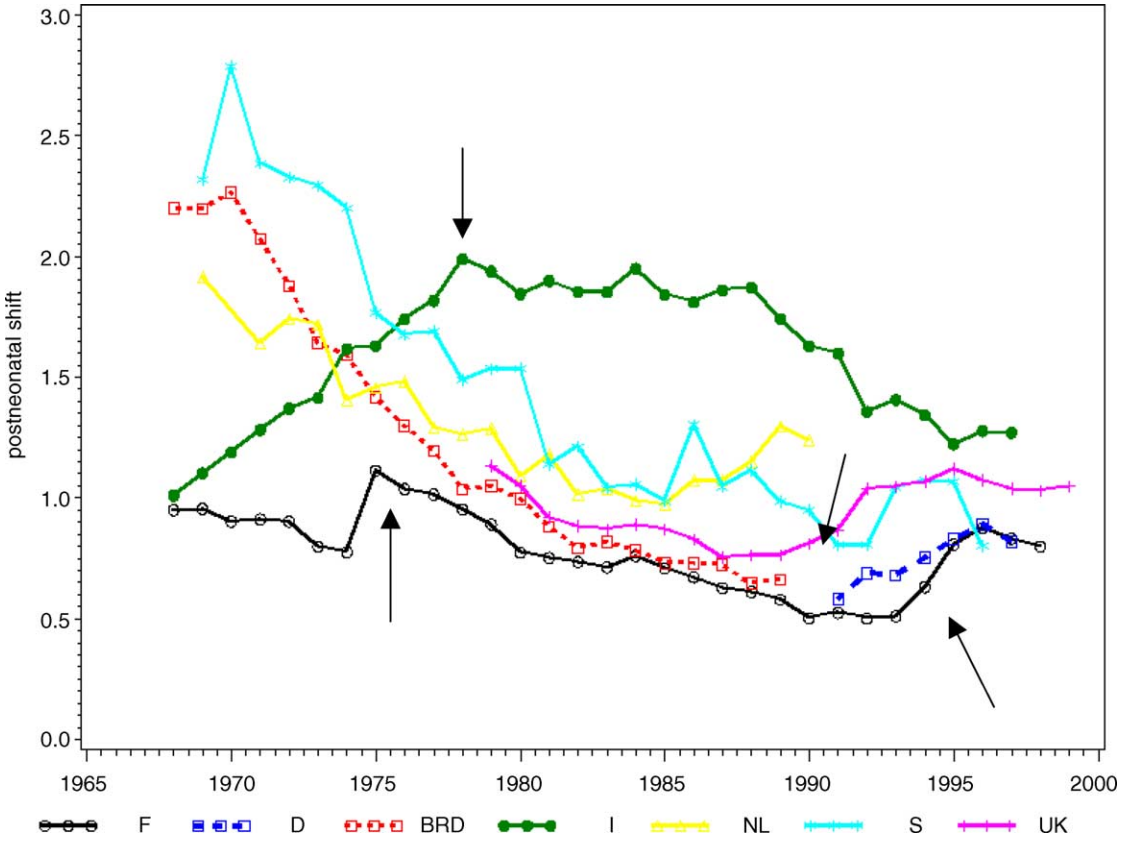

Fig. 2. Shift in late neonatal mortality for France, Germany before (BRD) and after (D) reunification, Italy, The Netherlands, Sweden and United Kingdom. Ratio of early neonatal (0–7 days) to late neonatal (8–28 days) deaths plotted by year for selected European countries. The overall pattern shows downward trends indicating that mortality shifted from the early neonatal to the late neonatal period. Jumps and trend reversals are presumably due to changes in national registration practices. Source: WHO 2002.

assumes that ovulation and fertilization occur 14 days after the first day of the LMP.

The magnitude of these differences is substantial. In a simulation on Canadian data for which both LMP and ultrasound measures were available, Blondel et al found that the preterm birth rate varied from 7.6% with last menstrual period to 9.1% with ultrasound estimates alone [4]. The impact of other decision rules was also simulated: use of LMP but modification for a discrepancy with ultrasound of 2 weeks (preterm birth rate = 7.8%), 10 days (8.1%), 7 days (8.5%) or 3 days (9.0%). Similar results have been reported elsewhere [5].

Nonetheless there is no gold standard for measuring the duration of gestation. Systematic use of ultrasound has been criticized for introducing biases—for instance, if female fetuses are already smaller than male fetuses when the ultrasound measures are done, then adjusting gestational age according to biometric measures introduces a bias, as confirmed by findings that female fetuses are more likely to be postterm than male fetuses when ultrasound is used to date pregnancies [6].

The use of ultrasound measures for establishing gestational age may vary between countries. It is possible that differences introduce a random measurement error between countries (if individual physicians have their own opinions about the relative value of ultrasound versus last menstrual period within each country), but it is more likely that country-specific practices induce systematic errors. These might be linked to rules for computing the ‘‘best estimate of the due date’’, to the availability of ultrasounds or timing of first prenatal visit, and possibly other factors such as preferred birth control methods. Research is necessary on these issues before we can be completely confident that this indicator is truly comparable.

2.3. Defining common measures of normal birth weight for Europe

The proportion of low birth weight babies (i.e. with a birth weight less than 2500 g) is one of the most frequently used indicators of perinatal health. Low birth weight is associated with higher mortality, morbidity and long-term impairment. The percentage of babies with a birth weight under 2500 g is one possible indicator of the proportion of high-risk babies in the population. When presented by gestational age, the distribution of birth weight provides one method of measuring growth restriction. This indicator is appealing because of its availability (almost all babies are weighed) and its

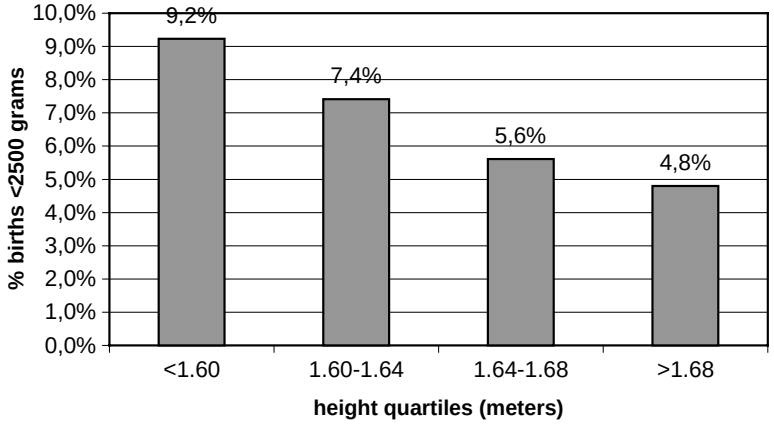

Fig. 3. Percentage of births <2500 by height quartile, French perinatal survey. *Source*: data from secondary analysis of French Perinatal Survey, 1998 [13].

reliability (errors in measurement are small and there is no reason to believe that scales give different results, on average, by hospital or country).

However, normal birth weights of babies in a population are affected by maternal and paternal factors, including height and weight, but also age and parity [7]. This shift in the overall distribution of birth weight associated with these biological characteristics affects the proportion of babies with low birth weight, when this is defined in absolute terms. An important fraction of babies with a low birth weight are preterm or growth-restricted or both, but some are normal-weight term babies. The proportion of the group of small and normal babies is likely to be larger among smaller women. Fig. 3 illustrates this association in a French sample: the proportion of babies weighing less than 2500 g decreases with each quartile of increasing maternal height.

The height of populations varies widely in Europe: in Portugal, Italy, Spain and Great Britain, the average height for women is between 1.61 and 1.63 m, whereas in Denmark, The Netherlands and Germany, it is between 1.67 and 1.69 m [8]. In countries where the average height is smaller, the proportion of normal and small babies included in the group with low birth weight would be expected to be higher. This appears to be borne out by Fig. 4, which plots the relation between the proportion of low birth weight babies in the population and the average height of women in the population (from EUROSTAT data) [8]. Although the proportion of low birth weight babies in the population is clearly determined by many factors, countries with shorter women tend to have a higher proportion of low birth weight babies.

One recent study of birth weight and mortality reports that the 'optimal' birth weight (defined as the birth weight at which mortality is lowest) varies in different European populations [9]. Further research is necessary to develop measures of appropriate birth weight that can be adapted to each country and used on a European level.

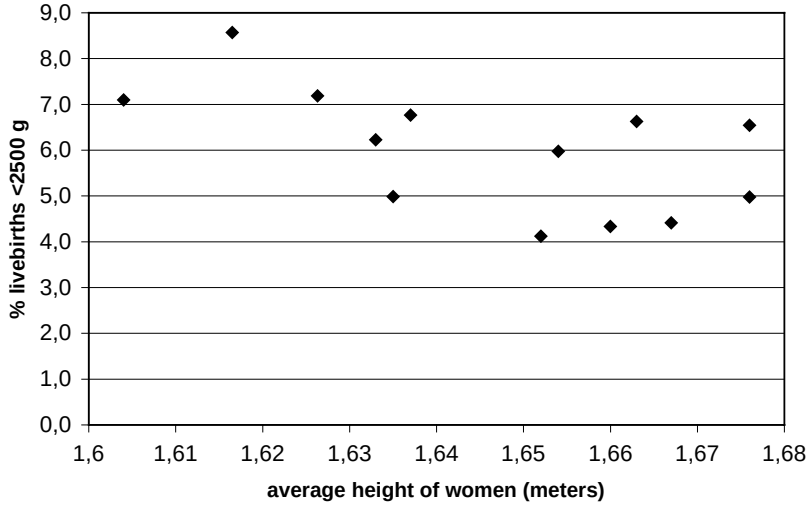

Fig. 4. Relation between women's average height and percentage of low birth weight live births. *Source*: PERISTAT data (see Buitendijk et al., this issue) and EUROSTAT (2000).

2.4. Coding mortality and morbidity

2.4.1. Changes in versions of the international classification of diseases (ICD)

As its name implies, the object of the ICD is to facilitate cross-country comparisons, but revisions and minor alterations may complicate this. Germany, for example, switched from ICD9 to ICD10 in 2000 for its coding for hospital in-patients. The substantial changes in the codes and their meanings between ICD9 and ICD10 implies a considerable shift in the spectrum of recorded diseases, and any cross-country comparison or longitudinal analysis will have to take this into account. The total number of codes has approximately doubled between the 9th and 10th revisions—from about 7000 to about 14,000 diagnosis codes [10,11]. In the ICD10 catalogue, chapters ''O'', ''P'' and ''Q'' are relevant to perinatology. Chapter ''O'' refers to pregnancy, delivery and the puerperium, chapter ''P'' details all conditions that have their origin in the perinatal period, and chapter ''Q'' lists congenital anomalies, deformities and chromosomal anomalies. An analysis of these codes shows that 163 ICD9 codes are mapped onto 235 ICD10 codes in chapter P and 180 ICD9 codes for anomalies onto 620 ICD10 codes. Changes in the ICD version used to register causes of death or morbidity will consequently result in systematic shifts in the overall levels reported. Tables comparing causes of death and morbidity across countries should therefore explicitly state the ICD version used for coding.

Some countries have taken steps to homogeneous coding practices on a national level. For instance, the Danish society of gynaecology and obstetrics has elaborated a guideline for registration of births which selects a number of codes from ICD10 and the Nordic classification of Surgical procedures and Treatments that were found to be relevant for registration on national level, with additional definitions and criteria for use where necessary [12].

2.4.2. The dangers of using data designed for other purposes than epidemiological surveillance: ICD and coding diagnosis-related groups (DRG)

Further biases are likely to arise when purposes of coding are altered, for example, when the classification of diseases is required for other than purely medical purposes. This is the case for the derivation of diagnosis-related groups (DRG). Examples from Germany illustrate this. The regulations governing remuneration of in-patient care are currently being changed from a calculation based on duration of hospital stay to a unit rate based on DRG. It is generally expected that this change will help to slow the increase in health care costs by providing incentives to reduce the average length of admissions. Because the monetary value of a DRG is also related to the patient's general morbidity, it is natural to expect that the number of coded ICD10 diagnoses, that is, the amount of comorbidity, will also rise and increase the complexity of the DRG.

Example 1: Perineal tears from vaginal delivery, grades 1 through 4, are coded O70.0, O70.1, O70.2 and O70.3, respectively. When associated with a standard vaginal delivery coded 9-260 in the International Classification of Procedures in Medicine (ICPM) the resulting DRG in the German system (G-DRG) will jump from O60D to O60C with a corresponding jump in the relative cost weight from 0.540 to 0.862 when moving from a second-degree tear O70.1 to a third-degree tear O70.2.

Example 2: A live born child (ICD10 code Z38.0) born without any further morbidity and receiving standard care (ICPM 9-262.1) is coded in DRG P67D with a relative cost weight of 0.436; for a ventricular septal anomaly (ICD10 Q21.3), the code is P67B and the relative cost weight increases to 1.397; with spina bifida (ICD Q05.5) as well, the code rises to P67A and the relative cost weight to 2.079.

Example 3: Ventricular haemorrhage (P10.2) results in increasing the relative cost weight from 0.476 to 1.397; if respiratory distress syndrome (P22.0) is also coded as present, the relative cost weight goes up to 2.079.

All three examples illustrate that once a DRG system is introduced for accounting purposes, it is likely that the levels of ICD10 diagnoses will change. This may affect the rates derived for maternal morbidity, causes of perinatal death and congenital anomalies.

2.5. Differences due to precision

Not all indicators are equally precise, that is, they may have quite dissimilar variance. Typically, rates computed for rare events have more variance, and this will increase further if it is based on small denominators. Apart from the fact that this is a direct consequence of the application of the mathematical formulas for computing variances and standard deviations, it means that observed differences in rates between countries will tend to be more significant when associated with lower variance and higher population denominators.

Within Europe, the units of comparison, that is, the countries vary greatly in size. This will ultimately affect the confidence intervals, especially for the smaller countries. Table 2 calculates the expected sampling variation for a rate of 1%, defined as the interval in which we would expect to find the estimated indicator 95% of the time if the underlying rate were 1%. The table demonstrates that the estimate will vary much more in smaller countries, even when the underlying rate is the same.

This effect is illustrated in Fig. 5, which reports regional perinatal mortality rates for Germany. The effect of sample size on confidence intervals may be seen from a plot of perinatal mortality rate per 1000 live and stillbirths against the number of births for 16 German regions (Bundesländer) for the year 1999. The data come from the German perinatal surveys (Fig. 8). The figure includes the upper and lower confidence intervals for the average perinatal mortality rate

Table 2
Country size and expected 95% sampling variation for a true population rate of 1%

Country name (coverage)	Data source	Number of total births	Lower 95% interval	Higher 95% interval
Austria	A1-2001	75707	0.77	1.23
Belgium	B1-1995	116122	0.82	1.18
Belgium (Flanders)	B2-2000	62122	0.75	1.25
Belgium (French community)	B3-2000	44328	0.71	1.29
Denmark	DK1-2000	66240	0.76	1.24
Finland	FIN1-2000	56768	0.74	1.26
France (perinatal survey)	F1-1998	13718	0.47	1.53
France	F2-2000	778341	0.93	1.07
Germany (nine Bundesländer)	D1-2000	558079	0.92	1.08
Germany	D2-1999	770744	0.93	1.07
Greece (perinatal survey)	EL1-1998	14659	0.49	1.51
Ireland	IR1-1999	54302	0.73	1.27
Ireland	IR2-1999	54242	0.73	1.27
Italy	I1-1998	533808	0.92	1.08
Luxembourg	L3-2000	5723	0.18	1.82
The Netherlands	NL-1999	201600	0.86	1.14
Portugal	P1-1999	120871	0.82	1.18
Spain	E1-1999	397632	0.90	1.10
Sweden	S1-2000	89722	0.79	1.21
UK	UK1,2,3-2000	681861	0.92	1.08

Note: data sources in Appendix A, described in Macfarlane et al. (this issue).

for Germany across all Bundesländer. The 95% confidence interval has a spread of at least 2 per 1000 for populations with fewer than 24,000 births per annum and a spread of at least 1 per 1000 for populations of fewer than 100,000 per annum. These spreads are considerable, as we see when we look at them in relation to the absolute size of the perinatal mortality rate: they correspond to approximately 40 and 20% of the perinatal mortality rate.

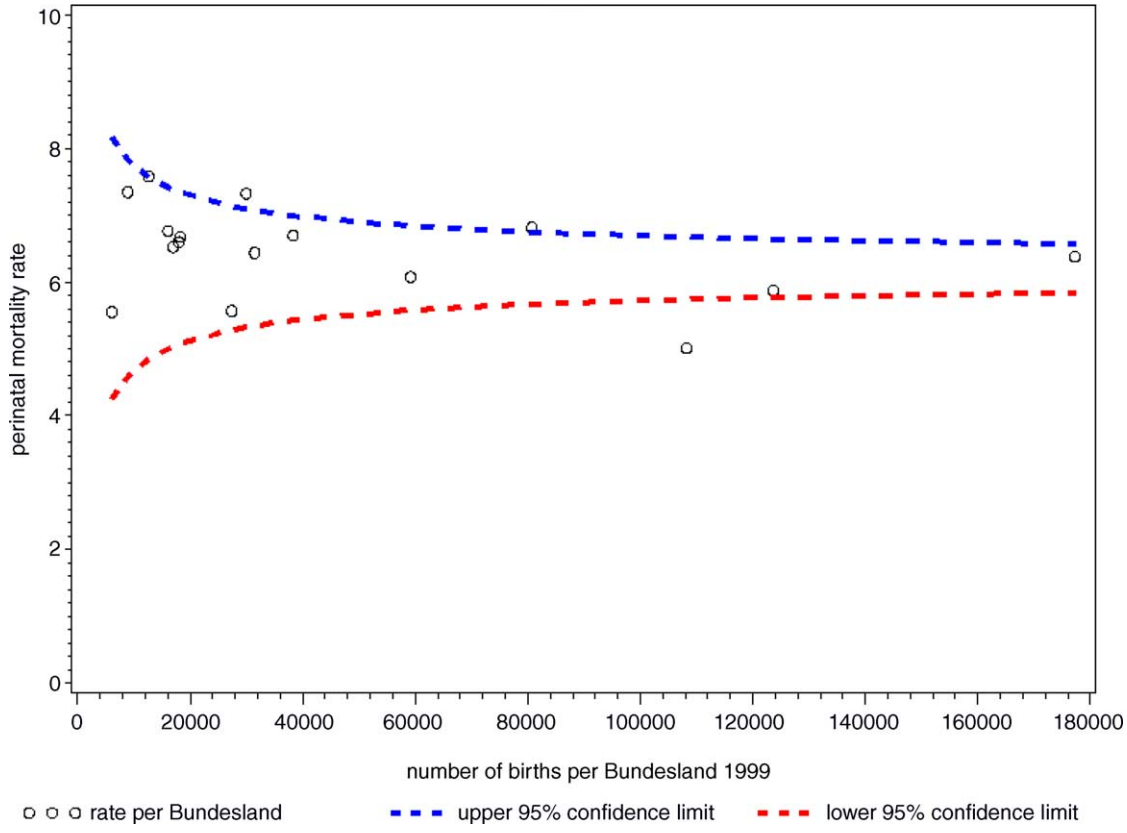

Fig. 5. Perinatal mortality rates per 1000 live- and stillbirths and 95% confidence intervals for the 16 German Bundesländer 1999. Spread is larger for smaller regions. *Source*: German perinatal surveys.

3. Conclusions and practical recommendations

The recommendations of the PERISTAT scientific advisory committee were made in light of these methodological difficulties in cross-European comparisons. The SAC aimed to minimize these methodological problems for the set of indicators it established, both through its selection process and its recommendations about the computation and presentation of indicators.

3.1. Choice of indicators

One of the key principles established by PERISTAT was to improve the quality of existing indicators, by implementing common definitions, data collection procedures and methods for constructing and presenting indicators. The group focused on developing valid indicators based on existing data sources, before suggesting new data for collection. The PERISTAT core set of indicators includes indicators that are relatively 'robust', on the grounds that it is better to have unbiased and comparable indicators that may not be extremely specific or sensitive, than indicators that are (theoretically) specific or sensitive, but will be unreliable and measured with bias at the European level.

The importance of developing truly usable indicators was a central tenet of the PERISTAT discussions and is evident in its results. The PERISTAT group incorporated this principle into the framework used for classifying the indicator set, by clearly distinguishing between indicators that can be used now and those that were desirable but require further work.

Consequently, not all issues are covered in equal density in the core PERISTAT indicator set. Maternal health and health care service related measures, for instance, are not equally represented at the core indicator level. Indicators to measure the intensity and quality of antenatal care provided for women are clearly needed. Post-delivery follow-up data must be collected to enable better assessment of the longer-term consequences of birth complications, such as perineal tears and faecal incontinence. Similarly, long-term assessment of neonatal health is inadequately covered. It is important to realise that the present indicator set allows only the assessment of a subset of perinatal health issues viewed through a narrow window centred in time on the moment of delivery. In other words, there is a trade-off between selection criteria for indicators, such as validity and comparability, versus sensitivity and comprehensiveness.

3.2. Computation and presentation of indicators

The PERISTAT project also addressed methodological problems in its recommendations on the computation and presentation of indicators. In particular, some indicators are computed by key subcategories to improve their interpretation and comparability (see Zeitlin et al., this issue). These subcategories are an integral part of the indicator definition. For instance, fetal and neonatal mortality should be presented by gestational age or birth weight groups. Buitendijk et al. (this volume) show how analysis by subgroups improves the interpretation and reliability of these data, by making it possible to separate out the groups, such as extremely low birth weight babies, for which comparability between countries is questionable.

Such straightforward stratification will often help to make comparisons more meaningful. The same is true for standardisation, and explains why the PERISTAT group recommends that indicators are often to be collected ''by'' other variables, such as maternal death by age. However, it must be remembered that all the comments regarding the quality of the indicators apply also to these other variables, and that any adjustment can only be valid if this new variable is collected homogeneously over Europe. The PERISTAT project has also defined indicators, whenever possible, as full distributions to improve our understanding of the variation in indicators between countries.

Because the way that data are presented can strongly influence the user's interpretation of their meaning, the PERISTAT project has emphasized the importance of providing guidelines for presentation and including confidence intervals and sample sizes, when necessary. Fig. 6a and b, which present the same data on neonatal mortality at term, illustrate the importance of this. Fig. 6a ranks countries according to mortality rate and uses a scale that accentuates even relatively minor differences. The graph calls attention to the disparity between the extreme high and low values. This graph contrasts sharply with Fig. 6b in which countries are ordered alphabetically, 95% confidence intervals are provided for the point estimates and the scale includes 0. This figure emphasises the uncertainty in the point estimates. Because of the small number of deaths at term in each country (and in particular in some countries, such as Luxembourg), precision varies substantially for this indicator. The user is more likely to conclude from this figure that the mortality rates at term are similar between the countries of Europe.

While the use of confidence intervals is important to underline that the countries of Europe differ substantially in the size of their populations, confidence intervals do not solve all problems. Systematic bias in registration and definitions will not be reflected in a measure of the statistical precision. In the example presented here, neonatal mortality of term births, we are confident that deaths and births are recorded. As explained above, this would not be the case for a comparison of neonatal mortality rates of births under 28 weeks of gestation.

3.3. Interpretation of indicators

Finally, the PERISTAT project recommends that a health reporting methodology be developed for presenting these indicators to users. Health reporting relies on specialists to

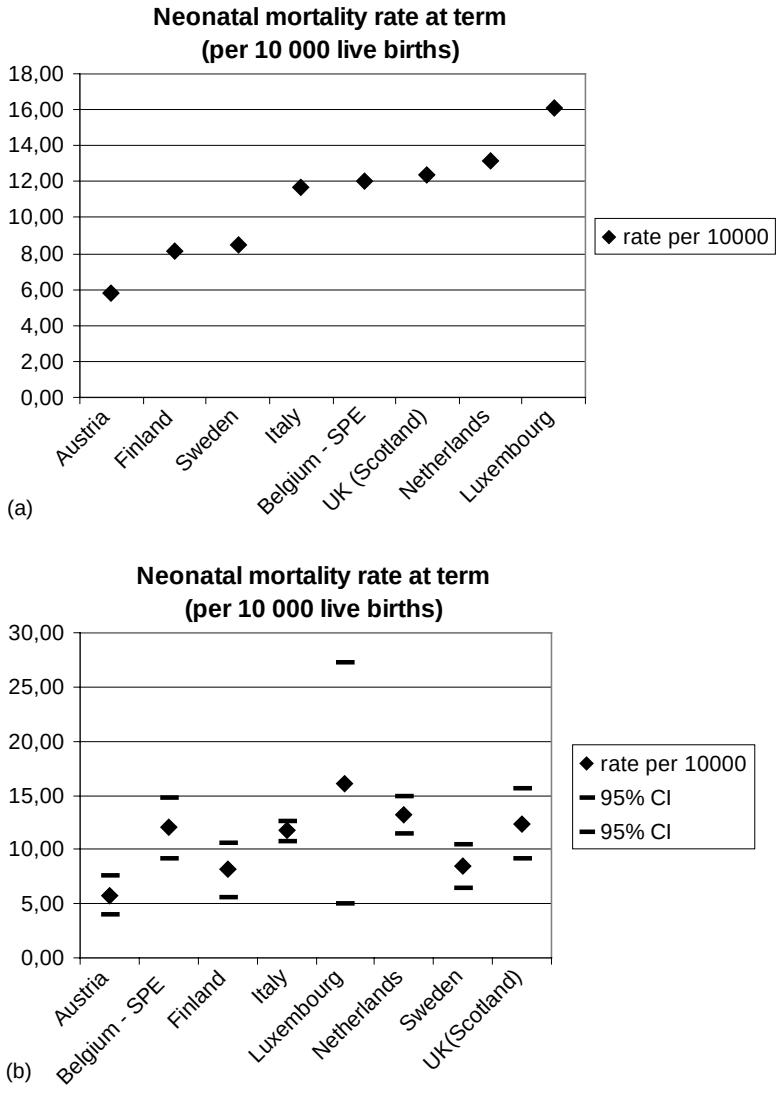

Fig. 6. (a and b) Neonatal mortality at term, per 10 000 births.

interpret information and provide 'meta data' on policy context that helps the non-specialist user to understand and appreciate the meaning of observed variations. Specialists can compare indicators with each other for internal validity and are familiar with the literature from perinatal health research, which can be used to confirm external validity.

Interpretation of indicators becomes easier as a health information system develops because the data are available over time; short-term fluctuations of indicators are more difficult to interpret than longer term trends. Furthermore, this historical perspective enriches health reporting by providing an additional dimension for understanding the development over time of health and health care in Europe as a whole and comparing them between individual countries.

Finally, enlisting perinatal health professionals, including clinicians, epidemiologists, statisticians and planners, to assist with health reporting on a national and European level should improve the interface between health surveillance and monitoring and health research and policy and further the general objectives of a health monitoring system.

Appendix A

Member state	Coverage (if not national)	Data source[a]	Year(s) provided	Abbreviation	Total births
Austria		Statistics Austria	2001	A1-2001	75707
Belgium					
		National Institute of Statistics and Scientific Institute of Public Health	1995	B1-1995	116122
	Flanders	SPE (Studiecentrum voor Pernatale Epidemiologie)	2000	B2-2000	62122
	French community	ONE (Office de la Naissance et de l'Enfance)	2000	B3-2000	44328
Denmark		Danish perinatal database	2000	DK1-2000	67337
Finland					
		Medical birth registry—STAKES	2000	FIN1-2000	56768
		Population statistics—Statistics Finland	2000	FIN5-2000	
France					
	Representative sample	National perinatal survey	1998	F1-1998	13718
		INSEE	2000	F2-2000	778341
Germany					
	Nine Bundesländer[b]	BAQ—perinatal survey	2000	D1-2000	558079
		Federal bureau of statistics Wiesbaden	1999	D2-1999	770744
Ireland		National Perinatal Reporting System	1999	IR1-1999	54302
Italy		ISTAT, Civil birth and death registration. Discontinued in 1998	1998	I-1998	533808
Luxembourg					
		FIMENA 2000	2000	L2-2000	5430
		Annuaire Statistique 2001	2001	L3-2001	5723
		Breast-feeding survey	2001	L4-2001	600
Netherlands		Merged database from professional registers. LVR: data on course of pregnancy and delivery; LNR: diagnoses of the child, duration of hospital stay, treatments	1999	NL-1999	201600
Portugal		Estatisticas Demograficas Estatisticas de Saude INE, Instituto Nacional de Estatistica	1999	P1-1999	120871
Spain					
		National Institute for Statistics (INE)	1999	E1-1999	397632 (live births)
	Valencia	GEN (Valencian group for neonatal studies)	2000	E5-2000	33467
	Valencia	General Direction of Public Health	2000	E6-2000	33467
Sweden		Medical Birth Register	2000	S1-2000	89722
UK					
	England and Wales	Office for National Statistics. Civil registration	2000	UK1-2000	607644
	Scotland	General Register Office, Scotland. Civil registration	2000	UK2-2000	53076
	Northern Ireland	General Register Office, northern Ireland. Civil registration	2000	UK3-2000	21512
	England	Department of Health, Maternity Hospital Episode Statistics	2000/01	UK4-00/01	

Appendix A. (*Continued*)

Member state	Coverage (if not national)	Data source[a]	Year(s) provided	Abbreviation	Total births
	Scotland	Information and Statistics Division, SMR2 Maternity Discharge Sheet	2000	UK6-2000	52413
	Northern Ireland	Perinatal Information, Northern Ireland, aggregated data from child health systems	2000	UK7-2000	21794
	Survey	Infant feeding 2000. Department of Health, the Scottish Executive, The National Assembly for Wales and the Department of Health, Social Services and Public Safety in northern Ireland	2000	UK15-2000	

[a] More detail on data sources provided in Mcfarlane et al. (this issue).

[b] Representing 72.6% of all births. Bayern, Baden-Württemberg, Berlin, Hessen (data from 2001), Niedersachsen & Bremen, Nordrhein, Sachsen, Thüringen, Westfalen-Lippe.

References

[1] World Health Organization. http://www.who.int/country/en, 2000.

[2] Graafmans WC, Richardus JH, Macfarlane A, Rebagliato M, Blondel B, Verloove-Vanhorick SP, et al. Comparability of published perinatal mortality rates in Western Europe: the quantitative impact of differences in gestational age and birthweight criteria. Bjog 2001;108:1237–45.

[3] Yang H, Kramer MS, Platt RW, Blondel B, Breart G, Morin I, et al. How does early ultrasound scan estimation of gestational age lead to higher rates of preterm birth? Am J Obstet Gynecol 2002;186:433–7.

[4] Blondel B, Morin I, Platt RW, Kramer MS, Usher R, Breart G. Algorithms for combining menstrual and ultrasound estimates of gestational age: consequences for rates of preterm and postterm birth. Bjog 2002;109:718–20.

[5] Savitz DA, Terry Jr. JW, Dole N, Thorp Jr. JM, Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. Am J Obstet Gynecol 2002;187:1660–6.

[6] Henriksen TB, Wilcox AJ, Hedegaard M, Secher NJ. Bias in studies of preterm and postterm delivery due to ultrasound assessment of gestational age. Epidemiology 1995;6:533–7.

[7] Gardosi J. Customized growth curves. Clin Obstet Gynecol 1997;40:715–22.

[8] EUROSTAT. Key Data on Health 2000. Luxembourg: European Commission; 2000.

[9] Graafmans WC, Richardus JH, Borsboom GJ, Bakketeig L, Langhoff-Roos J, Bergsjo P, et al. Birth weight and perinatal mortality: a comparison of ''optimal'' birth weight in seven western European countries. Epidemiology 2002;13:569–74.

[10] World Health Organization. International statistical classification of diseases, ninth revision. Clinical Modification. Annotated ICD9CM. Geneva: WHO; 1992.

[11] World Health Organization. International statistical classification of diseases and related health problems, 10th revision, vol. 1. Geneva: WHO; 1992.

[12] Danish Society of Obstetrics of Gynecology and Obstetrics. www.dsog.dk (in Danish); 2003.

[13] Foix-L'Helias L, BB. Changes in risk factors of preterm delivery in France between 1981 and 1995. Paediatr Perinat Epidemiol 2000;14:314–23.