A common data model for harmonization in the Nordic Pregnancy Drug Safety Studies (NorPreSS)

Jacqueline M. Cohen[1,2], Carolyn E. Cesta[3], Lars Kjerpeseth[1], Maarit K. Leinonen[4], Óskar Hálfdánarson[5], Øystein Karlstad[1], Pär Karlsson[3], Morten Andersen[6], Kari Furu[1,2] and Vidar Hjellvik[1]

1) Department of Chronic Diseases and Ageing, Norwegian Institute of Public Health
2) Centre for Fertility and Health, Norwegian Institute of Public Health
3) Centre for Pharmacoepidemiology, Karolinska Institutet
4) Information Services Department, Finnish Institute for Health and Welfare
5) Centre for Public Health Sciences, University of Iceland
6) Pharmacovigilance Research Center, Department of Drug Design and Pharmacology, University of Copenhagen
E-mail: jacqueline.cohen@fhi.no

ABSTRACT

It is necessary to carry out large observational studies to generate robust evidence about the safety of drugs used during pregnancy. In the Nordic countries, nationwide population-based health registers that document all births and dispensed prescribed drugs are valuable resources for such studies. A common data model (CDM) is a data harmonization and structuring tool that enables a unified and streamlined analytic approach for studies including data from multiple countries or databases. We describe a CDM developed for the Nordic Pregnancy Drug Safety Studies (NorPreSS), including details on data sources and structure of the data tables. We also provide an overview of the advantages and disadvantages of the approach (e.g. sharing of data analysis programs versus extra initial work to create CDM datasets from raw data).

I. INTRODUCTION

Studies of drug safety in pregnancy typically involve both rare exposures and outcomes. This necessitates conducting very large studies to generate robust evidence. In the Nordic countries, nationwide population-based health registers that include all births, dispensed prescribed drugs, and diagnoses from specialist care are valuable resources for such studies. International or multi-database studies that aim to have a single protocol and analytic approach must account for source data heterogeneity. One way of doing this is by data harmonization. In this paper, we describe the common data model (CDM) approach to data harmonization that the Nordic Pregnancy Drug Safety Studies (NorPreSS) consortium, which started in 2017, carried out to facilitate studies based on data from five Nordic countries. In part I we explain what a CDM is and the rationale for using one. In part II we describe the data sources used in NorPreSS. In part III we describe how the NorPreSS CDM is designed and populated, by giving the structure of the various data tables. In part IV we sum up with pros and cons of the CDM approach.

What is a Common Data Model?

In a workshop report from a 2017 meeting at the European Medicines Agency, a CDM was defined as, "…a mechanism by which raw data are standardised to a common structure, format and terminology independently from any particular study in order to allow a combined analysis across several databases/datasets" (EMA 2018). Gini et al. (2020) called this "a general CDM". A general CDM is made prior to and independent of any study protocol. The US FDA Sentinel CDM (https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model/) (Platt et al., 2018) and the Observational and Medical Outcomes Partnerships (OMOP) CDM (https://www.ohdsi.org/data-standardization/the-common-data-model/) (Kent et al., 2020) are examples of general CDMs. Both of these CDMs can be applied to pregnancy research (Matcho et al., 2018; Sentinel Operations Center 2019; https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model/mother-infant-linkage-table).

Our NorPreSS CDM is rather general but constructed with specific objectives in mind (to investigate immediate and long-term outcomes of drug use in pregnancy). It is a set of specific definitions for the structure of databases and data elements (basic units of information) that specifies translation of existing raw data elements in existing data into identically structured tables.

Designing, populating, and applying the CDM

The workflow from local raw data files via CDM datasets to final study results in NorPreSS is depicted in Figure 1. After designing the CDM, based on which data sources and variables are available in the different countries and necessary for the aims of the collaboration, the first step is to transform the country-specific data into

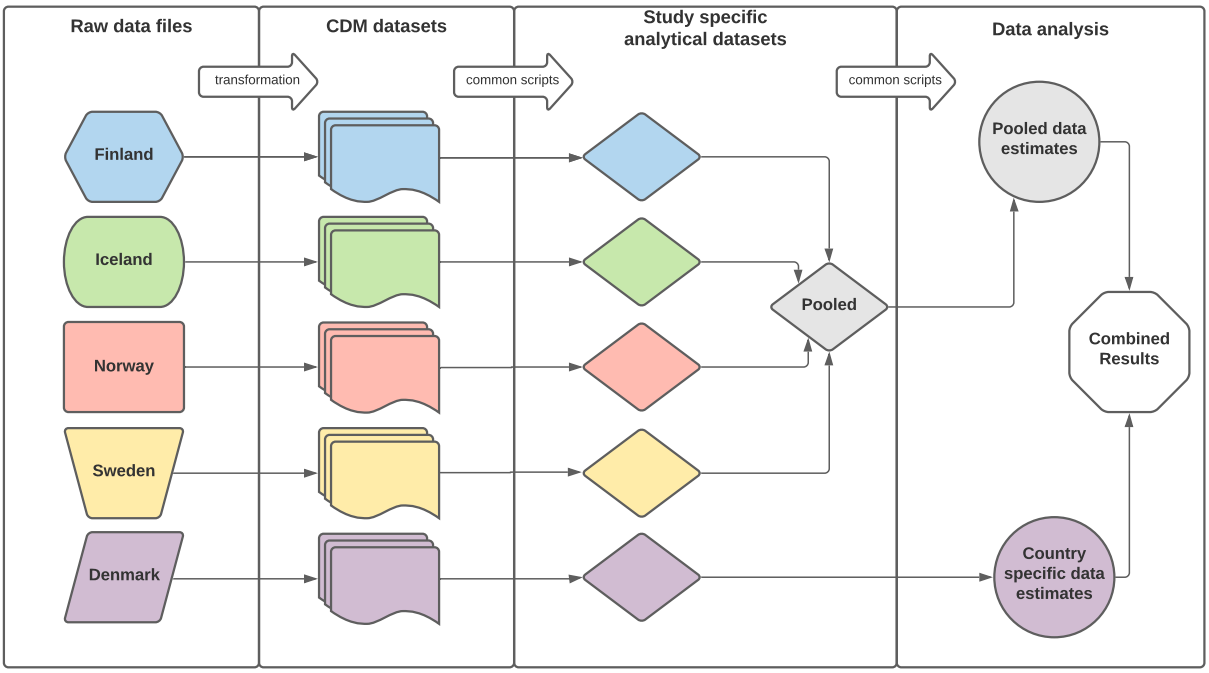

Figure 1. Workflow from original Nordic data to study results via a common data model (CDM). Creating a CDM facilitates both distributed data analysis and pooling of individual-level data.

the CDM format and populate the CDM – i.e. create the actual CDM datasets. The next step is to apply the CDM to specific research questions by creating study specific datasets according to study protocols. This typically involves mapping from CDM-variables to concepts like exposure, outcome, confounder, and is sometimes aided by "concept dictionaries" (or preconfigured rules systems) (Schneeweiss et al., 2020). A drug concept dictionary could for example be a look-up table containing all drug codes that represent an exposure, outcome, or confounder. We have not applied concept dictionaries in the NorPreSS project due to the inherent homogeneity of coding systems in the Nordic data. Further, due to the diverse study aims within our collaboration, we chose to leave these concept specifications to each study. When the CDM datasets are transferred to a central location, they can be analyzed together. When it is not possible to combine the data in a central location, common analysis scripts can be distributed to generate country-specific estimates that are later combined for a final study result.

Rationale for creating a CDM

Transforming data from different countries or healthcare providers into a uniform data structure allows for harmonized protocols, shared analysis programs, common result templates and pooling of individual-level data (in cases where data are allowed to cross national borders). In the NorPreSS project, the Finnish, Icelandic, Norwegian and Swedish data are transformed locally according to the CDM and are transferred and stored at the Norwegian Institute of Public Health where only authorized persons who work on the project can access the data. This facilitates quality checks by allowing side-by-side comparison of the contents of the populated CDM datasets from the four countries, pooling of the individual-level data, and reducing analytic personnel time since one programmer can analyze data from four countries. Data transfer to Norway requires data transfer agreements between the local data controller in Finland, Iceland and Sweden and the central data processor (Norway/the Norwegian Institute of Public Health). The Danish data are kept at Statistics Denmark due to legal regulations and analyzed by Danish researchers. A common protocol and CDM yield uniform results tables that facilitate combining results by meta-analytic techniques where aggregated data from different sites can be combined in a central place (Selmer et al., 2016). The NorPreSS CDM was developed by collaborators from Finland, Iceland, Norway, and Sweden, and later applied to the Danish data.

II. DATA SOURCES

Similar health and social registers exist in Denmark, Finland, Iceland, Norway, and Sweden, with some important differences, including variable names. In order to increase data privacy, some of the register holders provide the time since a reference date instead of actual calendar dates, for example the Norwegian linked register data when studies include the Norwegian Prescription Database. As such, actual dates for the other countries were converted to reference dates. Data from Finland was based on the Drugs and Pregnancy project, an ongoing study and infrastructure for drug safety studies in Finland with regular data updates for new

study years from the register holders (https://thl.fi/en/web/thlfi-en/research-and-expertwork/projects-and-programmes/drugs-and-pregnancy).

Medical birth registers (MBR)

Medical birth registers (MBRs) contain information on maternal, pregnancy, birth, and infant characteristics (Langhoff-Roos et al., 2014). All countries include live-births and stillbirths from 22 weeks with a unique record for each child. The MBR in Norway additionally includes pregnancies from 12 weeks including late pregnancy terminations which require approval by a special medical assessment board. In Finland, terminations of pregnancy for fetal anomaly (TOPFAs) are available from the Register of Induced Abortions and the Register of Congenital Malformations.

Several maternal conditions (e.g., asthma, epilepsy, diabetes) are captured in the MBR in Norway in a series of binary variables based on check boxes from standard antenatal care forms. There is also space to record ICD codes for other conditions. The MBRs in Iceland and Sweden include ICD codes for maternal conditions. Since 2004, the Finnish MBR also records some maternal conditions. However, for the entire study period, information on several conditions including epilepsy and asthma were appended from the Special Refund Entitlement Register and the Care Register for Health Care in Finland.

Infant outcomes, particularly major congenital anomalies, are an important focus of NorPreSS studies and are recorded in the MBRs. Norway's MBR includes major anomalies identified up to one year after birth with specific ICD-10 diagnoses as part of a complex string variable which indicates the source of information, code, and coding system. Iceland's MBR typically includes congenital anomalies diagnosed during the delivery hospitalization, and Sweden's MBR includes anomalies diagnosed within the first three months of life. Finland has a unique Register of Congenital Anomalies with validated diagnoses, classified according to the Atlanta/CDC modification of ICD-9 for classification of major congenital anomalies (ICD-9A). In Denmark, infant outcomes and records of TOPFAs are identified from the Danish National Patient Register (Bliddal et al. 2018, Broe et al. 2020).

Prescribed drug registers (PDR)

Prescribed drug registers (PDRs) include all prescribed drugs dispensed from pharmacies (Furu et al., 2010). In Finland during the study period, this only included drugs eligible for reimbursement. However, in future, it may be possible to consider all prescriptions and dispensations, regardless of reimbursement status (Aarnio et al. 2019). Each PDR includes a record for each drug product dispensed including the dispensing date, ATC code, strength, quantity, and amount of defined daily doses (DDDs) dispensed or a Nordic Article Number to obtain this information. None include prescribed dose in a structured format. The Norwegian Prescription Database (NorPD) includes the indication for reimbursement of reimbursed prescriptions. Indications for drug reimbursement are also tied to specific prescriptions in Finland, but only when the dispensed drug has been reimbursed in a special reimbursement category for chronic illness. Codes for which indication the drugs are prescribed for are not available in the PDRs in Denmark, Iceland, or Sweden.

Primary and specialized healthcare registers

Each country has a national patient register (NPR) that includes visits to specialist healthcare services and hospitals. They record admission and discharge dates for inpatient stays, the main diagnosis for each hospitalization, and other secondary diagnoses. Outpatient visits in NPRs include the visit date and any diagnoses recorded in association with that visit. Diagnoses are given as ICD-10 codes and procedures as Nordic Medico-Statistical Committee (NOMESCO) Classification of Surgical Procedures (NCSP) codes. Unique to Denmark, outpatient diagnoses are tied to a contact that may span a long time period and cover several visits. In Iceland, dates were only given as month and year, to reduce the risk of identifying individual study subjects. However, admission length was provided for inpatients. In Iceland, data was also obtained from the Centre for Child Development and Behavior and the State Diagnostic and Counselling Centre; two outpatient specialist registers on child neurodevelopment.

Primary care registers were also available and included in the CDM for Norway (available from 2006 and more complete from 2008), Iceland, and Finland (available from 2011 and with more complete coverage from 2013). They include International Classification of Primary Care (ICPC) codes and some ICD codes (and only ICD codes in Iceland) and visit date.

Cause of death registers and national statistics

Cause of death registers record the death date and underlying and contributing causes of death. The National Statistical agencies in each country, e.g. Statistics Norway, collect similar information on migration in and out of the country, socioeconomic position indicators including educational attainment, national academic assessments, and sick leave. In Iceland, academic assessment data were provided by the Directorate of Education, rather than Statistics Iceland. In Finland and Sweden, sick leave data were provided by the Social Insurance agencies. We included those data in our CDM to define the study population or censor individuals, adjust for confounding, or as outcomes, e.g. child academic performance, in specific NorPreSS studies.

III. DESIGNING THE CDM

The NorPreSS CDM is based on a CDM developed in the Cancer Risk and Insulin Analogues (CARING) project (But et al., 2017). The CARING CDM had three basic datasets: A study population dataset (one record

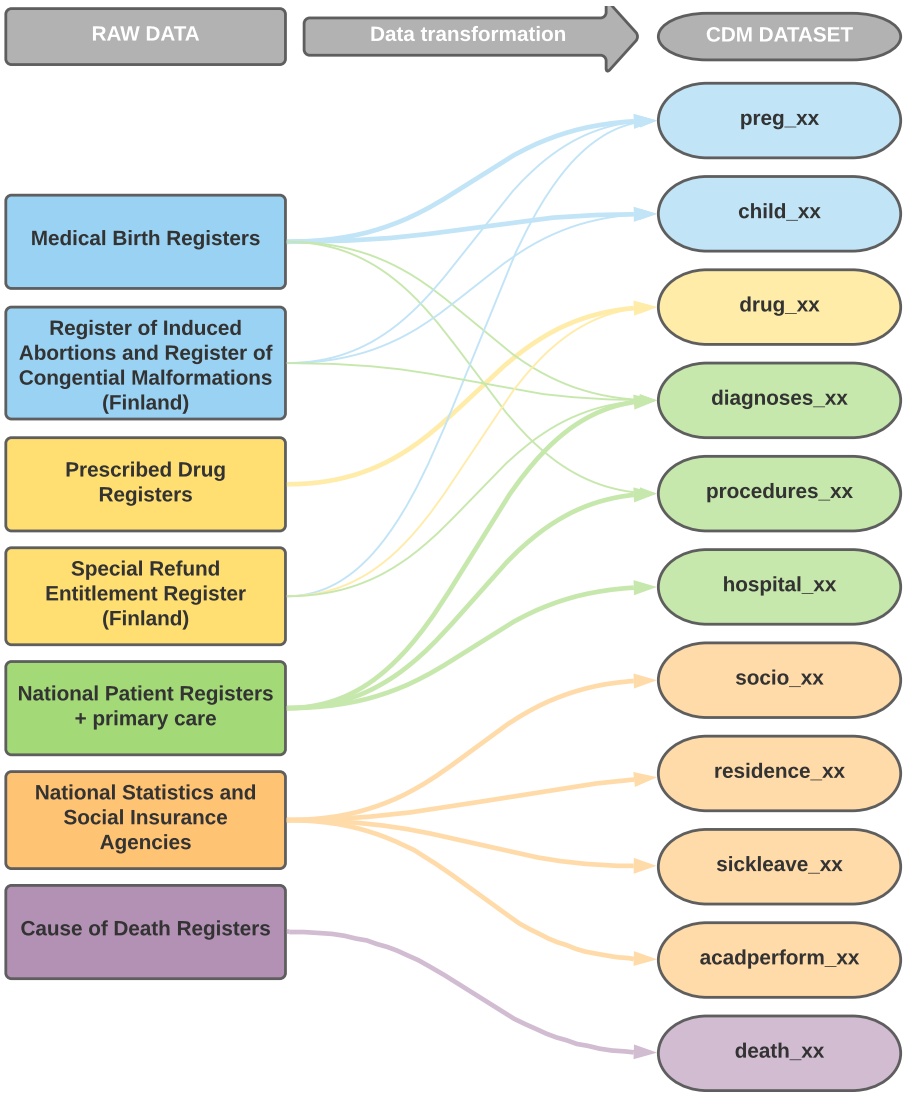

Figure 2. Overview of the NorPreSS Common Data Model. The original data sources are transformed into a set of identically structured tables for each country according to the NorPreSS Common Data Model (CDM). Bold lines indicate the main data source for each derived CDM dataset. We also have tables preg_xx_TOPFA and child_xx_TOPFA for pregnancy terminations when available (not shown).

per individual, with fixed person characteristics), a drug exposure dataset (one record per dispensed drug) and a clinical event dataset (one record per diagnosis or procedure). In NorPreSS, the unit of analysis is typically the pregnancy, requiring a pregnancy dataset linking each pregnancy to a mother and a child dataset linking each child/fetus to a pregnancy.

Researchers in each country create a set of identically structured data tables (Figure 2). Each file name ends in a 2-letter code indicating the country (_xx). File formats for the data tables are CSV files that can be imported into any statistical software. The full CDM specification is provided as Supplementary Information.

Unique identifiers

Since the study population focuses on pregnancies rather than individuals, we have several personal identifiers for the unique pregnancy (preg_id), mother (mother_id), and child (child_id). Information about the father/partner was not strictly necessary for our collaborative studies and was thus not available for every country. The preg_id variable was provided by the MBR in Iceland but is created by the researchers in other countries by grouping MBR records into distinct pregnancies. The mother and child IDs are encrypted IDs originally based on the personal identification numbers (PIN) assigned to all persons within the Nordic countries at birth or immigration. Therefore, child IDs have to be generated by the researchers in some instances to include pregnancies ending in stillbirth or abortion in our studies, where no PIN is registered.

Pregnancy data

Maternal and pregnancy characteristics are primarily identified from the MBRs and are collected into the table preg_xx which includes the unique pregnancy ID (preg_id) and mother ID (mother_id). The main

Table 1. The diagnoses_xx data table of the NorPreSS CDM.

Variable name	Variable label	Type	Valid values	Description
source_country	Source Country	CHARACTER	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person Identification	NUMERIC	Integer	10-digit number starting with: 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
hosp_no	Hospital Encounter Number	NUMERIC	Integer	See *person_id*
diag_type	Type of Diagnosis	NUMERIC	1/2	1=main diagnosis, 2=secondary diagnosis, or diagnoses made in primary care
diag_code_sys	Diagnosis Code System	CHARACTER	ICD10, ICPC2, ICD9A, SVA3	Diagnosis code system (including version when applicable)
diag_code	Diagnosis Code	CHARACTER		
diag_date	Date of diagnosis	NUMERIC	Integer	Admission, visit, or delivery date, number of days from a reference date
diag_source	Source of diagnosis	NUMERIC	1/2/3/4/5/6	1=inpatient care 2=outpatient care 3=primary care 4=birth register 5=register of congenital anomalies 6=special reimbursement register

preg_xx table excludes terminations since this data was not available for all countries. Instead, a separate preg_xx_TOPFA table is created. A separate table, child_xx, contains the child characteristics with a unique child ID (child_id), as well as the preg_id, and mother_id to allow for linkage between the preg_xx and child_xx tables.

Event data

The event data are derived from drug dispensing records in PDRs, diagnoses and procedures in MBR, NPR, and primary care data, cause of death registers, and national statistics data. For the tables including event data for both mother and child, a generic personal identifier (person_id) which corresponds to the mother_id or child_id in the preg_xx and child_xx tables is used. If a person born in the study period also gives birth in the study period, the same person_id will occur both as a child_id and a mother_id. With future data management in mind, we split each dataset with one for children and another for mothers with none or very few individuals appearing in both datasets (e.g. drug_child_xx and drug_mother_xx). With the exception of cause of death, these tables may contain multiple records per person_id.

The table drug_xx contains one record for each drug dispensed in the PDR. We include variables available in every PDR and additionally, variables for indications from Norway and Finland. We also include a variable to indicate the version of the ATC system that was used in the dataset. This is typically the year the data were delivered from the PDR. The WHO Collaborating Centre for Drug Statistics methodology takes a conservative approach with infrequent changes to existing ATC codes and DDDs. However, awareness of these changes and their potential impacts is important. With ATC version in the drug dataset, ATC codes and DDDs can be standardized to the most current system if future updates are appended to existing data. For example, since the Finnish data are part of the ongoing Drugs and Pregnancy project, different versions of the ATC coding system exist in this dataset, which is regularly updated with new cross-sectional data extraction appended to existing project data.

The table hosp_xx contains one record for each hospital or specialist care contact, both inpatient and outpatient in the NPR. It provides a way to efficiently count the number of hospitalizations or outpatient specialist visits in a period of time.

We gather the diagnoses from multiple data sources into one data table, diagnoses_xx (Table 1). The date assigned to a diagnosis code varies according to the source of the diagnosis. For NPRs, the date corresponds to either the outpatient visit date or inpatient admission date. For Denmark, we create a record for an outpatient diagnosis for every visit within a contact, not just the first or last visit date. For MBRs and the Finnish Register of Congenital Malformations, we assign the delivery date (or procedure date for terminations) and for the Special Refund Entitlement Register the start date of the special reimbursement entitlement as date of diagnosis. For primary care, it is the visit date. We include an identifier for whether the diagnosis was the main diagnosis for an inpatient admission or other diagnosis. Data for all procedures are in a table, procedures_xx, which is

prepared using almost the same method to assign dates, but including a precise procedure date when available.

We create a dataset containing one record for each residence period per person in the source country, residence_xx, based on the national population register. It contains a unique ID for each residence period, person_id, a start and end date for the period, and the type of end: either end of data extraction, emigration, or death.

The table death_xx is based on the cause of death register with one record per person_id. All countries could provide a main cause of death (recorded in the variable death_diag) and death date and some had additional contributing causes with potentially several diagnoses recorded in the same variable, death_othdiag.

Other data

For our planned studies, we defined several data tables for specific outcomes or covariates. A table for collecting information on maternal socioeconomic characteristics contains one record for each mother and year of delivery. We have so far only included highest level of education completed. Finally, to evaluate sick leave after pregnancy and child academic performance as study outcomes, tables for these data were tailored to a common way we could structure such data in all countries.

IV. CONCLUSIONS

Harmonizing data in a CDM facilitates international collaboration in studies where data contain roughly the same information but are organized differently in the various countries/databases. A CDM can be somewhere between completely general or custom-built for one specific study protocol. The NorPreSS CDM is rather general but constructed with specific objectives in mind. Those objectives were to investigate immediate and long-term outcomes of drug use and discontinuation in pregnancy (e.g. congenital anomalies, neurodevelopmental disorders, maternal sick leave). Although extra work is required to create the harmonized CDM datasets from local raw data, this is outweighed by advantages at later stages, including:

- Fewer resources are needed for data management and analysis since analysis scripts developed by one researcher can be applied to all countries' data
- Facilitates an increased knowledge of data before the analysis phase
- Improves transparency and consistency in data management and analysis
- Easy to expand analyses, e.g. to do sensitivity analyses, and address new research questions that fit within the approved uses of the data

The recent studies carried out within our collaboration before we had individually pooled and harmonized data were based on sharing aggregated data tables (Cesta et al., 2019; Reutfors et al., 2020, Cohen et al., 2020). This approach was labor intensive, requiring the analysis to be run in each country with a unique program adapted to the local data structure. We had to define the frequency dataset needed for all analyses up front, which limited our ability to change definitions, categorizations and perform sensitivity analyses. Our current approach is more streamlined and flexible. We have the possibility to develop the analysis scripts based on pooled data from four countries. Combining two instead of five datasets in the final step (Figure 1) reduces the chances of having zero or very few outcomes among the exposed in one of them. Thus, for rare events, pooling individual data is preferable both for statistical and data privacy reasons (Selmer et al., 2016).

One should, however, be aware that the process of transforming data from local raw data to CDM datasets may introduce errors that are not so easily detected. A comparison of the different CDM variables by country and other quality checks should be done to look for anomalies. The translation process may also result in large files and computing capacity issues. Another potential disadvantage is losing some granular information that is available in only some countries. For example, if age or education is grouped in broad categories in one or more countries, the CDM adopts that broad categorization for all countries. Replacing exact calendar dates with reference dates creates challenges in interpreting data on an absolute time scale. We thus built in variables that anchor the observations in calendar time, according to year/quarter of birth. Further, even though the indication for a specific prescription was only fully available in the Norwegian Prescription Database, we accommodated this in the CDM since it was valuable for our studies.

The NorPreSS CDM is being used for multiple ongoing studies in progress and under peer review (Cohen et al., 2019; Cesta et al., 2020a; Halfdanarson et al., 2020; Kjerpeseth et al., 2020). It was used in one study published in 2020 assessing the risk of major anomalies in children with prenatal exposure to modafinil, a drug primarily indicated for narcolepsy (Cesta et al., 2020b). The study, based on data from two countries, went from concept to publication in 5 months, demonstrating our ability to rapidly assess pressing drug safety questions once the data have been harmonized in a CDM, developed for such a purpose.

Funding

The study was partly supported by NordForsk Nordic Program on Health and Welfare (Nordic Pregnancy Drug Safety Studies, project No. 83539), by the Research Council of Norway (International Pregnancy Drug Safety Studies, project No. 273366) and by the Research Council of Norway through its Centres of Excellence funding scheme (project No. 262700).

REFERENCES

Aarnio E, Huupponen R, Martikainen JE, Korhonen MJ. First insight to the Finnish nationwide electronic prescription database as a data source for pharmacoepidemiology research. *Res Social Adm Pharm* 2020;**16**:553-9.

Bliddal M, Broe A, Pottegård A, Olsen J, Langhoff-Roos J. The Danish Medical Birth Register. *Eur J Epidemiol* 2018;**33**:27-36.

Broe A, Damkier P, Pottegård A, Hallas J, Bliddal M. Congenital Malformations in Denmark: Considerations for the Use of Danish Health Care Registries. *Clin Epidemiol* 2020;**12**:1371-80.

But A, De Bruin ML, Bazelier MT, Hjellvik V, Andersen M, Auvinen A, et al. Cancer risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. *Diabetologia* 2017;**60**(9): 1691-1703.

Cesta CE, Cohen JM, Pazzagli L, Bateman BT, Bröms G, Einarsdóttir K, et al. Antidiabetic medication use during pregnancy: an international utilization study. *BMJ Open Diabetes Res Care* 2019;**7**(1):e000759.

Cesta CE, Halfdanarson OO, Cohen JM, Einarsdottir K, Furu K, Gissler M, et al. SSRI/SNRI discontinuation in pregnancy and obstetrical outcomes: A Nordic population register-based study [5116]. *Pharmacoepidemiol Drug Saf* 2020a;**29**(S3):557.

Cesta CE, Engeland A, Karlsson P, Kieler H, Reutfors J, Furu K. Incidence of Malformations After Early Pregnancy Exposure to Modafinil in Sweden and Norway. *JAMA* 2020b;**324**(9):895-897.

Cohen JM, Leinonen MK, Alvestad S, Bjørk MH, Cesta CE, Einarsdóttir K, et al. Comparative safety of antiepileptic drugs and risk of major congenital malformations. *Pharmacoepidemiol Drug Saf* 2019; 28(S2):13.

Cohen JM, Cesta CE, Furu K, Einarsdóttir K, Gissler M, Havard A, et al. Prevalence trends and individual patterns of antiepileptic drug use in pregnancy 2006-2016: A study in the five Nordic countries, United States, and Australia. *Pharmacoepidemiol Drug Saf* 2020;**29**(8):913-922.

EMA. A Common Data Model for Europe? – Why? Which? How? Workshop report from a meeting held at the European Medicines Agency, London, United Kingdom, 11-12 December 2017 [EMA/614680/2018]. October 4, 2018.

Furu K, Wettermark B, Andersen M, Martikainen JE, Almarsdottir AB, Sørensen HT. The Nordic countries as a cohort for pharmacoepidemiological research. *Basic Clin Pharmacol Toxicol* 2010;**106**(2):86-94.

Gini R, Sturkenboom MCJ, Sultana J, Cave A, Landi A, Pacurariu A, et al, Working Group 3 of ENCePP. Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model. *Clin Pharmacol Ther* 2020;**108**(2):228-235.

Halfdanarson OO, Cohen JM, Karlstad Ø, Cesta CE, Leinonen MK, Ozturk B, et al. Antipsychotic use during pregnancy and neurodevelopmental outcomes in children: A Nordic population register-based study [82]. *Pharmacoepidemiol Drug Saf* 2020;**29**(S3):513.

Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: Evidence generation for health technology assessment. *Pharmacoeconomics* 2021;**39**(3):275-285.

Kjerpeseth LJ, Cesta CE, Engeland A, Furu K, Gissler M, Gulseth HL, et al. Risk of major congenital malformations with metformin compared with insulin in pregnancy [424]. *Diabetologia* 2020;**63**(S1):S209.

Langhoff-Roos J, Krebs L, Klungsøyr K, Bjarnadottir RI, Källen K, Tapper AM, et al. The Nordic medical birth registers–a potential goldmine for clinical research. *Acta Obstet Gynecol Scand* 2014;**93**(2):132-137.

Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. *PLoS One* 2018;**13**(2):e0192033.

Platt R, Brown JS, Robb M, McCellan M, Ball R, Nguyen MD, Sherman RE. The FDA Sentinel Initiative – An evolving national resource. *N Engl J Med* 2018;**379**(22):2091-2093.

Reutfors J, Cesta CE, Cohen JM, Bateman BT, Brauer R, Einarsdóttir K, et al. Antipsychotic drug use in pregnancy: A multinational study from ten countries. *Schizophr Res* 2020;**220**:106-115.

Schneeweiss S, Brown JS, Bate A, Trifirò G, Bartels DB. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. *Clin Pharmacol Ther* 2020; **107**(4):827-833.

Selmer R, Haglund B, Furu K, Andersen M, Nørgaard M, Zoëga H, Kieler H. Individual-based versus aggregate meta-analysis in multi-database studies of pregnancy outcomes: the Nordic example of selective serotonin reuptake inhibitors and venlafaxine in pregnancy. *Pharmacoepidemiol Drug Saf* 2016;**25**:1160-1169.

Sentinel Operation Center. Mother-infant linkage: frequently asked questions & appendices, Version 1.0.0, February 2019.

Supplementary Information. Full Specification of NorPreSS CDM Tables

Table S1. Pregnancy related variables, preg_xx. Dataset contains one record per pregnancy resulting in a birth. A separate file preg_xx_TOPFA contains one record for each pregnancy with a termination of pregnancy for fetal anomaly and includes at minimum source_country, preg_id, mother_id, deliv_date, mother_age_cat, and others as available.

Variable Name	Variable Label	Type	Valid values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
preg_id	Pregnancy ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
mother_id	Mother ID	NUM	Integer	Same as *preg_id*
deliv_date	End of pregnancy date	NUM	Integer	End of pregnancy date, with reference dates
birth_yr	Delivery year	NUM	Integer	Delivery year
mother_age_cat	Maternal Age in categories	NUM	0/1/2/3/4/5/6	Mother's age <20=0, 20-24=1, 25-29=2, 30-34=3, 35-39=4, 40-44=5, >=45=6
mother_birth_country_other	Mother foreign born	NUM	0/1	0=born inside the source country, 1=born outside source country
mother_birth_country_nonnordic	Mother not born in Nordic	NUM	0/1	0=born in DK, IS, FI, NO, SE, 1=born outside the Nordic countries
cohabit	Married or cohabitating	NUM	0/1	Married/cohabiting
parity	Parity	NUM	0/1/2/3/4	0,1,2,3,4+ previous births
prev_still_births	Previous still births	NUM	Integer	Number of previous still births
prev_live_births	Previous live births	NUM	Integer	Number of previous live births
multiple	Multiple pregnancy	NUM	0/1	1= the birth is a plural/multiple birth (i.e. twins, triplets, etc.)
multiple_n	Plurality	NUM	Integer	Plurality, number of children born
gest_age_wks	Gestational age in weeks	NUM	Integer	Gestational age, best clinical estimate in completed weeks
gest_age_days	Gestational age in days	NUM	Integer	Gestational age, best clinical estimate in days
chronic_asthma	Asthma	NUM	0/1	Asthma complicating pregnancy
chronic_htn	Pre-existing hypertension	NUM	0/1	Pre-existing hypertension
chronic_renal	Chronic renal disease	NUM	0/1	Chronic renal disease

chronic_epilepsy	Epilepsy	NUM	0/1	Epilepsy before pregnancy
chronic_diabetes	Pre-existing diabetes	NUM	0/1	Pre-existing diabetes, type 1 or type 2
gest_htn_only	Gestational hypertension	NUM	0/1	Pregnancy-induced hypertension only, without pre-eclampsia
preeclampsia	Preeclampsia	NUM	0/1	Any preeclampsia (mild or severe)
eclampsia	Eclampsia	NUM	0/1	Eclampsia
gest_diabetes	Gestational diabetes	NUM	0/1	Gestational diabetes
folate_before	Folate use before pregnancy	NUM	0/1	Folate supplementation before pregnancy
folate_during	Folate use during pregnancy	NUM	0/1	Folate supplementation during pregnancy
smoke_beg	Smoking in early pregnancy	NUM	0/1	Smoking in early pregnancy (yes, no)
smoke_beg_num	Number of cigarettes early	NUM	0/1/2	0=non-smoker 1=1-9 cig per day (light), 2=10+ per day (heavy)
smoke_end	Smoking at end of pregnancy	NUM	0/1	Smoking in late pregnancy (yes, no)
height	Maternal height	NUM	numerical	Maternal height in cm
weight	Maternal pre-pregnancy weight	NUM	numerical	Maternal weight before pregnancy in kg
bmi	Maternal pre-pregnancy BMI	NUM	numerical	Maternal BMI before pregnancy, recorded in MBR
spont_labour	Spontaneous labor	NUM	0/1	Spontaneous labor
delivery_mode	Mode of delivery	NUM	0/1/2/9	0=vaginal; 1=planned c-section; 2=emergency c-section; 9=unspecified c-section; missing = no information on mode of delivery
abruption	Placental abruption	NUM	0/1	Placental abruption
prom	Premature rupture of membranes	NUM	0/1	Premature rupture of membranes
art	ART pregnancy	NUM	0/1	Assisted reproductive technology treatment
birth_outcome	Birth outcome	NUM	0/1/2	0=all live births, 1=all still births, 2= both live and still births

Table S2. Child related variables, child_xx. Dataset contains one record for each born child. A separate file, child_xx_TOPFA, contains one record for each termination of pregnancy for fetal anomaly and includes at minimum source_country, preg_id, mother_id, child_id, gest_age_child, child_birth_yr, child_birth_qrt, and others as available.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
preg_id	Pregnancy ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
mother_id	Mother ID	NUM	Integer	Same as *preg_id*
child_id	Child ID	NUM	Integer	Same as *preg_id*
child_order	Birth order	NUM	Integer	Birth order in the pregnancy
child_birthday	Birth date	NUM	Integer	Reference day of birth
child_male	Male	NUM	0/1	0=female; 1=male
gest_age_child	Gestational age at birth	NUM	Integer	Gestational age at birth according to best clinical estimate
child_birth_yr	Birth year	NUM	Integer	Birth year
child_birth_month	Birth month	NUM	Integer	Birth month
child_birth_qrt	Birth quarter	NUM	Integer	Birth quarter / 1=Jan-Mar, 2=Apr-Jun, 3=Jul-Sept, 4=Oct-Dec
birth_weight	Birth weight	NUM	Integer	Birth weight in gram
birth_length	Birth length	NUM	Integer	Length in cm
head_circum	Head circumference	NUM	Integer	Head circumference in cm
apgar1	Apgar 1 minute	NUM	0-10	Apgar
apgar5	Apgar 5 minutes	NUM	0-10	Apgar
apgar10	Apgar 10 minutes	NUM	0-10	Apgar
nicu	Transfer for NICU	NUM	0/1	Transfer to neonatal intensive care unit, as recorded in MBR
stillborn	Stillborn	NUM	0/1	Stillborn (died before or during delivery)
neonatal_death	Neonatal death	NUM	0/1	As recorded in MBR, death in first 28 days (0-27 days)
breech	Breech presentation	NUM	0/1	1=breech presentation in pregnancy

Table S3. Dispensed prescription variables, drug_xx (drug_mother_xx, drug_child_xx). Dataset contains one record for each prescription dispensation event.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
disp_date	Dispensing date	NUM	Integer	Days since a reference date
disp_year	Dispensing year	NUM	Integer	Calendar year of dispensing
disp_qrt	Dispensing quarter	NUM	1/2/3/4	1=Jan-Mar, 2=Apr-Jun, 3=Jul-Sept, 4=Oct-Dec
atc_version	ATC version	NUM	Integer	Version year
atc_code	ATC code	CHAR	valid ATC Codes	One dispensed prescribed drug
pack_size	Quantity in the package	NUM		Number of tablets, g, ml etc per package
pack_units_tablets	Tablets as units	NUM	0/1	Tablets or other formulation
pack_ndispensed	N. of packages dispensed	NUM	Integer	Number of packages dispensed
total_tablets_dispensed	Total tablets dispensed	NUM	Integer	Total number of pills: tablets or capsules
total_ddd	Total DDDs	NUM		
strength	Medication strength	NUM		
strength_units	Strength units	CHAR		
drug_pack_id	Package ID	NUM		Product ID corresponds to the unique Nordic Article Number
prescriber_type	Type of prescriber	NUM	1/2/3/4/5	1=obstetrics, 2=psychiatry 3=neurology 4=primary care, 5=other
indication_ICD10	Indication Norway, ICD-10	CHAR		Norwegian reimbursement indication in ICD10 code (from 2008)
indication_ICPC	Indication Norway, ICPC	CHAR		Norwegian reimbursement indication in ICPC2 code (from 2008)
Indication_refnum	Indication Norway, refnum	CHAR		Norwegian reimbursement indication according to law paragraph (to 2008)
indication_FI	Indication Finland	CHAR		Finnish special reimbursement indication

Table S4. Hospital encounter variables, hosp_xx (hosp_mother_xx, hosp_child_xx). Dataset contains one record for each hospital contact, both inpatient and outpatient.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
hosp_no	Hospital encounter number	NUM	Integer	Sequential number for healthcare recorded encounters in dataset, not linked to *person_id*, but same format
adm_type	Admission type	NUM	1/2	1=inpatient care, 2=outpatient care
adm_plan	Planned admission	NUM	1/2	1=planned, 2=not planned
adm_date	Admission date	NUM	Integer	Admission date for either outpatient visit or inpatient hospitalization. Days since reference date
disc_date	Discharge date	NUM	Integer	Days since reference date. Missing for outpatient visits
prov_type	Type of provider	NUM	1/2/3/4/5	Type of provider. 1=obstetrics, 2=psychiatry 3=neurology 4=primary care, 5=other
adm_date_prec	Precision of admission date	NUM	1/2/3/4	1=day, 2=week, 3=month, 4=year
disc_date_prec	Precision of discharge date	NUM	1/2/3/4	1=day, 2=week, 3=month, 4=year. Only for inpatient care

Table S5. Diagnosis variables, diagnoses_xx (diagnoses_mother_xx, diagnoses_child_xx). Dataset contains one record per diagnosis.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number starting with: 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
hosp_no	Hospital encounter number	NUM	Integer	Sequential number for healthcare recorded encounters in dataset, not linked to *person_id*, but same format.
diag_type	Type of diagnosis	NUM	1/2	1=main diagnosis, 2= secondary diagnosis, or diagnoses made in primary care or other source
diag_code_sys	Diagnosis code system	CHAR	ICD10, ICPC2, ICD9A, SVA3	Diagnosis code system (including version when applicable)
diag_code	Diagnosis code	CHAR		
diag_date	Date of diagnosis	NUM	Integer	From admission, visit, or delivery date, relative to reference date
diag_source	Source of diagnosis	NUM	1/2/3/4/5/6	1=inpatient care, 2=outpatient care, 3=primary care, 4=birth register, 5=register of congenital anomalies, 6=special reimbursement register

Table S6. Procedure variables, procedures_xx (procedures_mother_xx, procedures_child_xx). Dataset contains one record for each procedure.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
hosp_no	Hospital encounter number	NUM	Integer	Sequential number for healthcare recorded encounters in dataset, not linked to *person_id*, but same format.
proc_code_sys	Procedure code system	CHAR		Procedure Code System (incl. Version when applicable), eg 'NCSP' or relevant
proc_code	Procedure code	CHAR		E.g. NOMESCO (NCSP-code)
proc_date	Date of procedure	NUM	Integer	Procedure date or admission date. Days since reference date
proc_source	Source of procedure data	NUM	1/2/3/4	1=inpatient care, 2=outpatient care, 3=primary care, 4=birth register

Table S7. Socioeconomic variables, socio_xx. Dataset contains mothers' education level in the year of birth for each pregnancy. It could be expanded to include other socioeconomic variables such as income.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
preg_id	Pregnancy ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
mother_id	Mother ID	NUM	Integer	Same as *preg_id*
mother_educ	Education	NUM	0/1/2/3	Education level in the year of delivery. 0=compulsory or less, 1=secondary, 2=post-secondary, 3= postgraduate

Table S8. Residence periods, residence_xx (residence_mother_xx, residence_child_xx). Dataset contains one record per residence period in the source country.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
res_period_id	Residence period	NUM	Integer	Individual-level ID: 1 for the first residence period per person, 2 for a second, etc.
res_period_start	Start date of residence period	NUM	Integer	According to reference dates
res_period_end	End date of residence period	NUM	Integer	According to reference dates
res_period_end_type	Residence period ending type	CHAR	S/E/D	S=End of data extraction, E=Emigration, D=Death

Table S9. Cause of death, death_xx (death_mother_xx, death_child_xx). Dataset contains one record for each death.

Variable Name	Variable Label	Type	Valid Values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
death_date	Date of death	NUM	Integer	Date of death. Days since reference date
death_diag	Underlying cause of death	CHAR		ICD code for diagnosis (i.e. cause of death)
death_othdiag	Contributing causes of death	CHAR		Semicolon-separated ICD codes
ddiag_vers	Death diagnosis version	CHAR		Version of Code classification, E.g. ICD-10
death_basis	Basis for identifying cause of death	NUM	1/2/3/4/9	1=Clinical autopsy, 2=Extended forensic, 3=Forensic Autopsy, 4=Forensic examination that does not include autopsy, 9=No autopsy or forensic examination
inj_poi	Injury or poisoning	NUM	1/2/3/4/5	1=accident, 2=intentional self-inflicted, 3=intentionally inflicted by another person, 4=unclear whether the intent existed, 5=other external cause of death

Table S10. Sick leave variables, sickleave_xx. Dataset contains one record per pregnancy, reflecting sick leave use before and during pregnancy, and in the 1.5 years following delivery. It is tailored to the research questions related to maternal sick leave within the NorPreSS collaboration.

Variable Name	Variable Label	Type	Valid values	Description
source_country	Source country	CHAR	DK/IS/FI/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
preg_id	Pregnancy ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 4=Finland, 5=Denmark
mother_id	Mother ID	NUM	Integer	Same as *preg_id*
sickleave_before	Sick leave before	NUM	0/1	Sick leave episode recorded between LMP-365 and LMP-1; 0=no, 1=yes
sickleave_during	Sick leave during	NUM	0/1	Sick leave start date recorded between LMP and delivery-1; 0=no, 1=yes
sickleave_after	Sick leave after	NUM	0/1	Sick leave start date between birthdate and delivery+547 days (1.5 years); 0=no, 1=yes
sickleave_after_days	Sick leave days after	NUM	0-547	Number of days of sick leave between birthdate and 1.5 years after delivery.
sickleave_after_date	Sick leave after date	NUM	integer	Start date of first sick leave episode after delivery. Days since reference date.
sickleave_before_indication	Sick leave before indication	CHAR		Indication for sick leave between LMP-365 and LMP. If there are multiple indications (ICD codes), include semicolon-separated list
sickleave_before_ind_psych	Sick leave before psychiatric indication	NUM	0/1	Psychiatric indication for sick leave between LMP-365 and LMP; 1=psychiatric diagnosis (including burnout) ICD-10 F10-F99, X60-X84 0=other
sickleave_during_indication	Sick leave during indication	CHAR		Indication for sick leave between LMP and delivery. If there are multiple

				indications (ICD codes), include semicolon-separated list
sickleave_during_ind_psych	Sick leave during psychiatric indication	NUM	0/1	Psychiatric indication for sick leave between LMP and delivery; 1=psychiatric diagnosis (including burnout) ICD 10 F10-F99, X60-X84 0=no psychiatric diagnosis
sickleave_after_indication	Sick leave after indication	CHAR		Indication between delivery and delivery+1.5 years. If there are multiple indications (ICD codes), include semicolon-separated list
sickleave_after_ind_psych	Sick leave after psychiatric indication	NUM	0/1	Psychiatric indication between delivery and delivery+1.5 years; 1=psychiatric diagnosis (including burnout) ICD 10 F10-F99, X60-X84 0=no psychiatric diagnosis
disability_before	Disability pension before	NUM	0/1	Disability pension between LMP-365 and LMP; 0=no, 1=yes
disability_after	Disability pension after	NUM	0/1	Disability pension between delivery and delivery+1.5 years; 0=no, 1=yes

Table S11. Academic performance variables, acadperform_xx. This dataset contains one record for each test subject.

Variable Name	Variable Label	Type	Valid values	Description
source_country	Source country	CHAR	DK/IS/NO/SE	DK=Denmark, IS=Iceland, FI=Finland, NO=Norway, SE=Sweden
person_id	Person ID	NUM	Integer	10-digit number - starting with 1=Iceland, 2=Norway, 3=Sweden, 5=Denmark
child_age_yr	Age of child in years	NUM	Integer	Child age in years at the time of testing
child_age_months	Age of child in months	NUM	Integer	Child age in months at the time of testing
child_grade	Grade level, school year	NUM	Integer	Grade that the child is in at the time of test
child_behind_ahead	Years child is behind or ahead own age group	NUM		Indicates whether a child is behind or level with their age group, e.g. -1=1 year behind, 0=level, 1=1 year ahead, etc.
test_yr	Year of test	NUM	Integer	Test year
test_month	Month of test	NUM	Integer	Test month
test_subject	Subject of test	CHAR	Language or Math	Test subject, language or math
test_participation	Participation in test	NUM	0/1	0=Non-participation, 1=Participation
test_exempted	Exempted from test	NUM	0/1	0=Not exempted, 1=Exempted
test_fail	Test result	NUM	0/1	0=Pass, 1=Fail
percentile_rank	Percentile rank	NUM	Integer	Score percentile based on rank of test score in test subject and year