ARTICLE

Data Quality in the Ontario Midwifery Program Database, 2006 to 2009

Adriana Cappelletti, BHSc, Angela H. Reitsma, RM, MSc, Julia Simioni, MSc, Jordyn Horne, BSc, Caroline McGregor, BSc, Rashid J. Ahmed, BSc, and Eileen K. Hutton, PhD

ABSTRACT

***Objective**: To identify common errors in midwifery data collection and provide midwives with the rationales behind data cleaning, the importance of reliable data, and the links between data collection, research studies, and evidence-based care.

Methods: A database containing records of all women who received midwifery care in Ontario that was invoiced to the Ministry of Health and Long-Term Care between April 1, 2006, and March 31, 2009, was obtained. Data cleaning was performed to assure that the data set was as complete and accurate as possible. Duplicate records were identified and removed. Missing, inconsistent, and implausible data were identified and corrected where possible or removed.

Results: Common data errors included inappropriate use of open text fields and drop-down menus, incorrect interpretation of "planned place of birth," reporting of outcomes that should be mutually exclusive, and reporting of incorrect, incomplete, or missing information.

***Discussion**:* Midwives have an important role in the collection of health information that is complete and accurate. Several common errors were identified that, if corrected, would improve the quality of midwifery data and in turn would contribute to high-quality research, which will inform midwifery practice, policy-makers, and women and their families about midwifery care.

KEYWORDS
midwifery, data collection, Ontario, database, data quality

This article has been peer reviewed.

ARTICLE

Qualité de la base de données sur le programme de formation des sages-femmes en Ontario, de 2006 à 2009

Adriana Cappelletti, BHSc, Angela H. Reitsma, s.-f. aut., MSc, Julia Simioni, MSc, Jordyn Horne, BSc, Caroline McGregor, BSc, Rashid J. Ahmed, BSc, et Eileen K. Hutton, PhD

RÉSUMÉ

Objectif : Identifier les erreurs courantes de collecte de données sur la profession de sage-femme. Fournir aux sages-femmes les motifs justifiant le nettoyage des données; insister sur l'importance de la fiabilité des données; établir des liens entre la collecte de données, les études de recherche et les soins fondés sur des données probantes.

Méthodes : Nous avons accédé à une base de données renfermant le dossier de toutes les femmes ayant reçu les soins d'une sage-femme en Ontario et dont le ministère de la Santé et des Soins de longue durée a reçu la facture entre le 1er avril 2006 et le 31 mars 2009. Nous avons nettoyé cet ensemble de données pour en assurer, autant que possible, l'exhaustivité et l'exactitude. À ce titre, nous avons identifié et éliminé les dossiers en double. Nous avons également identifié et corrigé les données manquantes, contradictoires et invraisemblables, dans la mesure du possible, ou nous les avons supprimées.

Résultats : Les erreurs fréquentes de collecte de données englobaient l'utilisation inadéquate des zones de texte ouvertes et des menus déroulants, l'interprétation incorrecte du « lieu prévu pour l'accouchement », le signalement de résultats qui devraient être mutuellement exclusifs et la transmission de renseignements inexacts, incomplets ou manquants.

Discussion : Les sages-femmes jouent un rôle important dans la collecte de renseignements exacts et complets sur la santé. Nous avons identifié plusieurs erreurs courantes dont l'élimination améliorerait la qualité des données sur la pratique sage-femme et, en retour, contribuerait à la réalisation de travaux de recherche de grande qualité. En plus d'enrichir la pratique des sages-femmes, de tels travaux de recherche informeront les décideurs, ainsi que les femmes et leur famille, sur les soins prodigués par ces professionnelles de la santé.

MOTS CLÉS

pratique sage-femme, collecte de données, Ontario, base de données, qualité des données

Cet article a été soumis à l'examen collégial.

BACKGROUND

Clinical studies of serious adverse maternal and infant outcomes are challenging due to the large sample sizes required to study rare outcomes. Obstetric interventions and practices are not always suitable for evaluation with randomised controlled trials, due to a lack of equipoise and to the unwillingness of potential participants to have their birth plans determined by randomisation.[1] The retrospective use of registries or administrative databases is cost-effective, avoids selection bias, and does not require participant recruitment.[2]

The success of retrospective analyses of databases is highly dependent on the quality of the database design and the quality of the data.[2,3] To ensure high quality, the data must be complete, correct, consistent, plausible, and up to date.[2] Data quality assurance includes error prevention, which is the reduction of errors at the time of data measurement, collection, and entry. Once the data are entered into the database, questionable data can be found by looking for values that are missing, outside of the expected range, or inconsistent with other data (logic checks). Once questionable data are identified, the data user must determine if the value should be altered (usually in accordance with other data fields for the case), removed, or left unchanged. This process of detecting and rectifying questionable data is termed "data cleaning."[4]

All midwives in Ontario are required to submit prenatal, intrapartum, and postpartum information for each client to the Ministry of Health and Long-Term Care (MOHLTC).[5] Since 2003, the submission of this data has been tied to reimbursement, which together with ongoing audit virtually ensures a complete database. Prior to 2012, the data were maintained by the Ontario Midwifery Program (OMP) within the MOHLTC. The OMP collected maternal and infant data by using a six-page "Ontario Maternal Newborn Health Form" (OMP form), which consisted of information on billing and services (p. 1); a record of a woman's antepartum, intrapartum, and postpartum conditions and care (pp. 2-4); and an infant record capturing details of the birth, neonatal conditions and care, and feeding (pp. 5–6).

Our research team previously conducted a retrospective cohort study using OMP data collected between 2003 and 2006 to compare maternal and perinatal/neonatal mortality and morbidity and intrapartum intervention rates for women who planned homebirth attended by midwives at the onset of labour with low-risk women who planned a hospital birth. This work resulted in an important and highly cited paper for Canadian midwifery.[5] To conduct further studies of midwifery care in Ontario, our research team obtained records of midwife-booked pregnancies from the MOHLTC that were invoiced between April 1, 2006, and March 31, 2009.

Through the data-cleaning process, our team discovered several areas on the OMP form that were commonly completed improperly. The purpose of this descriptive report is to identify common errors in midwifery data collection and to provide midwives with the rationales

Adriana Cappelletti, Angela H. Reitsma, Julia Simioni, Jordyn Horne, and Caroline McGregor *are associated with the Midwifery Education Program of the Faculty of Health Sciences at McMaster University in Hamilton, Ontario.*

Rashid J. Ahmed *is a member of the Department of Obstetrics and Gynecology at the Michael G. DeGroote School of Medicine at McMaster University in Hamilton, Ontario.*

Eileen K. Hutton *is associated with both the Midwifery Education Program of the Faculty of Health Sciences at McMaster University and the Department of Obstetrics and Gynecology at the Michael G. DeGroote School of Medicine at McMaster University.*

behind data cleaning, the importance of reliable data, and the links between data collection, research studies, and evidence-based care.

DATA-CLEANING METHODOLOGY

Data about (1) maternal and neonatal outcomes, (2) patterns of care received during the prenatal, intrapartum, and postpartum periods, (3) gestational age at significant time points, and (4) place of birth were obtained from the OMP at the MOHLTC with an agreement to allow analyses for the purpose of our proposed studies. Research ethics board approval was obtained as required by McMaster University. De-identified records of all women who received midwifery care in Ontario that was invoiced to the MOHLTC between April 1, 2006, and March 31, 2009, were received. The database did not contain Ontario Hospital Insurance Plan numbers and therefore could not be linked to any individual's health records.

Data cleaning was performed with SPSS Statistics 22.0. The purpose of this process was to ensure that the data set was as complete and accurate as possible without access to original client records for confirmation of questionable data. Duplicate records were detected by searching for records with an identical maternal date of birth, infant date of birth, birth weight, infant time of birth, and birth hospital postal code. In the case of identical duplicate records, only one record was retained; the other(s) were removed. Duplicate records that differed in respect to at least one data field were assessed by two experienced midwives to determine whether they were actually records of two pregnancies or an error. If the two reviewers could not reach a consensus, a third reviewer was involved.

After removing all likely duplicate records, we assessed the completeness of each variable. The proportion of missing data for each variable was the number of records that actually had a value among those that were expected to have a value. Missing and questionable data could not be found or verified against any source documents; thus, we imputed values by using other data fields wherever possible. Inconsistent and implausible data were identified by using logic checking to compare one data field with other data fields that contained redundant information and by comparing them to a set of legal data ranges that were developed with the intention of flagging outliers. Imputations were made to correct or remove unlikely and impossible values. The algorithms for the imputation of missing, outlying, and illogical values were created on a variable-to-variable basis. When necessary, two experienced midwives reviewed records to determine how the data should be handled. If they did not reach a consensus, a third midwife reviewer was involved.

FINDINGS

The data set received from the MOHLTC contained 54,249 records. After the removal of records that were likely to be duplicates, 54,026 records remained. The following sections offer descriptions of

Adriana Cappelletti, Angela H. Reitsma, Julia Simioni, Jordyn Horne et Caroline McGregor *sont affiliées au programme de formation des sages-femmes de la Faculté des sciences de la santé de l'Université McMaster, établie à Hamilton (Ontario).*

Rashid J. Ahmed *est membre du département d'obstétrique-gynécologie de l'École de médecine Michael G. DeGroote de l'Université McMaster, établie à Hamilton (Ontario).*

Eileen K. Hutton *est affiliée au programme de formation des sages-femmes de la Faculté des sciences de la santé de l'Université McMaster, ainsi qu'au département d'obstétrique-gynécologie de l'École de médecine Michael G. DeGroote, à l'Université McMaster.*

common data errors and provide examples from the OMP database.

Inappropriate Use of Open Text Fields and Drop-Down Menus

The OMP form contained several questions that were answered by choosing from a list of outcomes or interventions. The questions included both an option to indicate "other" and an open text box where an item that was not on the list could be reported. These open text boxes were commonly used either to report an item that was in fact on the list of options or to report an outcome or intervention that was not relevant to the question. For example, to record laceration taking place during the intrapartum period, the OMP form allowed midwives to record one or more of the following: first-degree perineal, second-degree perineal, third-degree perineal, fourth-degree perineal, cervical, labial, vaginal, other (with open text field), or none. We found that midwives commonly used the open text field to specify a type of laceration already listed on the form. Additionally, we found that midwives often misused the open text field, describing outcomes not considered to be laceration, including cesarean section (n = 141) and episiotomy (n = 11), among other entries such as details on suturing of the wound.

As another example, to record complications and other pregnancy- and birth-related conditions, midwives were to indicate as many as necessary from a list of 53 items (including "other") and had the option to use the open text box to indicate a condition that was not on the list. We found a high prevalence of complications selected as "other" for events that were specifically listed on the form.

Selecting the appropriate outcome on the checklist, rather than indicating "other," is important for research and other uses of the data. As a consequence of using "other" in the examples above, women who experienced the event may not have been "captured" for the condition or complication that they actually experienced. The inaccuracy resulting from this misuse of the "other" field leads to underestimation of the prevalence of the outcome of interest. By contrast, indicating "other" in order to use the open text box to write notes that were not relevant inflates the number of women who experienced the outcome in question and leads to overestimation of the prevalence of the condition.

Incorrect Interpretation of "Planned Place of Birth"

Planned place of birth is listed on the OMP form as "Began intrapartum period intending to give birth at," followed by the option to select "home," "hospital," "undecided," or "other." We performed logic checks as part of our comparative study of planned-home versus planned-hospital birth and concluded that several records in which home was indicated as the planned place of birth were not truly planned homebirths at the onset of labour, although this may have been the plan at some time during pregnancy. Examples of such inconsistencies with a planned homebirth included preterm births before 37 weeks' gestation (and as early 26 weeks' gestation) and breech births by cesarean section with no labour. The recording of these births as planned homebirths, perhaps due to plans made during the pregnancy before complications arose, will cause any adverse events among these higher-risk groups to be erroneously attributed to planned homebirths.

Reporting of Outcomes That Should Be Mutually Exclusive

One of the steps in the data-cleaning process was to rectify any contradicting data for each record. Owing to the lack of verifying documents, this was challenging and in many cases required two midwives to review the records in detail to deduce what likely happened, which was a time-consuming process.

In regard to stillbirth, live birth, and neonatal death, one question requested midwives to decide whether the birth was a live birth or a stillbirth; neonatal death was to be reported in answer to a subsequent question. In cleaning the data set, we found eight records indicating both a stillbirth and neonatal death (less than seven completed days), which are mutually exclusive events. Although this number seems small relative to the total number of records in the data set (n = 54,026), the low frequency of infant death in Ontario signifies the importance of distinguishing between these two events.

The reporting of two mutually exclusive events is problematic because it requires the data user to make a decision about which event presumably occurred. If not done carefully, this could unintentionally lead to possible misclassification bias. For example, the decision could be based on which outcome to choose, based on what interventions the case was exposed to (in the case of our study, whether the woman had planned a homebirth or a hospital birth).

Impossible Outlying Data

Data entry errors are common, and mistakes such as inversion of digits can lead to extreme outliers whose occurrence is virtually impossible. These types of errors occurred for all continuous and date variables and resulted in impossible values for maternal age, height, weight, and infant birth weights, as well as for infants born before the conception date. Incorrect admission and discharge dates also led to negative values for length of hospital stay. Impossible values were removed from the data set and could not be obtained from source documents, therefore increasing the amount of missing data and decreasing the number of records that could be included in analyses of those variables.

Incomplete or Missing Data

Most of the data in the OMP data set were complete. The percentage of missing data for each variable ranged from 0% to 49% of records, with a 0.19% median proportion of missing data. Some data fields, however, were frequently absent. For example, maternal weight at booking was missing 49% of the time. For example, under "Infant Discharge Date/Time," midwives are asked to record the infant discharge date and time if the infant was not discharged with his or her mother, in order to determine infant length of stay. We found that 12% (527 of 4,287) of infants who were admitted to the neonatal intensive-care unit did not have a discharge date, nor did 11% (216 of 2,055) of infants who were not discharged from the hospital with their mothers. These infants represent the group of infants with poorer health outcomes; therefore, it would be helpful for midwives to obtain discharge data on these infants even if the infants are no longer receiving midwifery care. Accurately calculating infant length of stay is important, for example, when describing the utilization of health services by midwifery clients or when conducting an economic analysis.

As another example, midwives may record infant feeding as breast milk only, no breast milk, or a combination of breast milk and other liquids or food. Data on infant feeding at birth and at three days postpartum were fairly complete; however, data on infant feeding at 10 days postpartum (or at the next closest midwife visit up to four weeks) were missing for 15% of records. Efforts to capture this information are important for assessing long-term breastfeeding activity among women receiving midwifery care.

In some cases, missing data were inferred if other information was available, but for the majority of records with missing data, no changes were made. In either case, bias is potentially introduced either through selection or through misclassification.

DISCUSSION

Midwives have an important role in providing accurate data that are critical to quality midwifery research and, one might argue, have a professional responsibility to provide such data. Data sets themselves are becoming more sophisticated with inbuilt logic checks, which will not allow mutually exclusive events to be entered. As data sets become more integrated, it will likely be possible to link data between midwifery data sets and, for example, hospital data sets, so that outcomes data such as discharge time and date for infants who are no longer in midwives' care can be accurately accessed. In the meantime, ensuring that data entry is complete and accurate will contribute to high quality research that will inform midwifery practice, policy makers, and women and their families about midwifery care. The Canadian model of practice is unique, and our clinical outcomes are of international interest.

In Ontario, the OMP data set has been taken over by Better Outcomes Registry and Network, making the data more accessible to researchers and allowing linkage with outcomes from other providers. Now more than ever, it is important that midwifery data reflect midwifery care.

ACKNOWLEDGEMENTS

We would like to thank the Ontario Midwifery Program at the Ministry of Health and Long-Term Care for providing the data that was analyzed, and the Association of Ontario Midwives for providing an open research grant that was used to fund part of our work with this data set.

REFERENCES

1. Hendrix M, Van Horck M, Moreta D, Nieman F, Nieuwenhuijze M, Severens J, et al. Why women do not accept randomisation for place of birth: feasibility of a RCT in the Netherlands. BJOG. 2009:37–42.
2. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.
3. Beretta L, Aldrovandi V, Grandi E, Citerio G, Stocchetti N. Improving the quality of data entry in a low-budget head injury database. Acta Neurochir (Wien). 2007 Jan;149(9):903–9.
4. Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005 Oct;2(10):e267.
5. Hutton EK, Reitsma AH, Kaufman K. Outcomes associated with planned home and planned hospital births in low-risk women attended by midwives in Ontario, Canada, 2003-2006: a retrospective cohort study. Birth. 2009 Sep;36(3):180–9.