

WASHINGTON STATE DEPARTMENT OF HEALTH

# Improved Birth/CHARS Linkage: CHINCHILLA vs. UW BERD



Maya Bhat, MPH & Sean Coffinger, MA



DOH 422-296 April 2026

To request this document in another format, call 1-800-525-0127. Deaf or hard of hearing customers, please call 711 (Washington Relay) or email [doh.information@doh.wa.gov](mailto:doh.information@doh.wa.gov).

# Table of Contents

Table of Contents .....	2
Acknowledgements .....	3
Background .....	3
BERD VS. CHINCHILLA .....	4
<b>Methods</b> .....	<b>4</b>
UW BERD .....	4
CHINCHILLA .....	4
<b>2014 Comparison</b> .....	<b>5</b>
<b>Limitations</b> .....	<b>5</b>
<b>Using CHINCHILLA for analysis</b> .....	<b>6</b>
Appendix A: CHINCHILLA Model Accuracy .....	7

## Acknowledgements

The LIDA (Linkage and Integrated Data Analysis) unit in CHS works collaboratively with several talented data linkage scientists, data scientists, data engineers and epidemiologists. Our work would not be possible without the contributions from the Vital Records staff and expertise provided by SME epidemiologists: Danielle Legeai, CHARS Epidemiologist; Kelly Matzke, Birth Epidemiologist; Sam Rolland, VR Epidemiology Manager. Support from the Data Science and Engineering unit and collaboration with other linkage scientists warrants acknowledgements: Julian Kapoor, Linkage Data Scientist.

# Background

The Center for Health Statistics (CHS) has made available an annual data set consisting of birth records linked to hospital discharge records from the Comprehensive Hospital Abstract Reporting System (CHARS). This file is the only routinely produced statewide data set that combines demographic, newborn health, prenatal care, and childbirth information from the birth certificate with ICD-CM diagnosis codes from CHARS hospital discharge records for both mother and infant. This data set has been used by epidemiologists and other analysts as an important and unique data source for reproductive health research projects.

Initially, from 1987 to 1998, the linked data set was produced by Washington State Department of Health and was known as 'Birth Events Records Database' (BERD). Then, in 1999, production of the file was taken over by University of Washington's Department of Epidemiology using an updated method for identifying links. Files created from 1999 to 2014 are referred to as 'UW BERD' files to distinguish them from other data sets created using different linkage methods. Finally, for birth years 2016 and onward, CHS resumed the creation of this data set using a new linkage method and plans to continue releasing them annually. Deviating from the BERD acronym, these newest files are known as 'CHINCHILLA' files: 'CHARS INfant CHILdbirth-events Linked Automatically. For each birth year there is one file for infants on the birth record linked to their own CHARS records (Baby CHINCHILLA) and a second file for mothers listed on the birth record linked to their own CHARS records (Mother CHINCHILLA).

The purpose of this document is to compare baby CHINCHILLA and UW BERD methods of linkage and offer recommendations for using the data for analysis. Detailed information regarding newer methods implemented by CHS can be found here: [Analytical Methods and Reports | Washington State Department of Health](#).

<b>BERD</b>	1987 - 1998
<b>UW BERD</b>	1999 - 2014
<b>Baby CHINCHILLA</b> <b>Mother CHINCHILLA</b>	2016 - Current

# BERD VS. CHINCHILLA

## Methods

### UW BERD

These files consisted of links between births (mother and newborn) of a given year and the corresponding CHARS records **for the birth event only**. CHARS records were restricted by gender and age (females between 10 and 69 years old) and by birth-related ICD-9 diagnosis codes (V3x.xx). Ideally, this resulted in a one-to-one link between the newborn on a birth record and his/her own CHARS record and similarly for the birth mother and her respective CHARS birth-event discharge. Both newborn and maternal links were contained in a single data file.

The UW BERD method relied on a set of criteria that were applied hierarchically to determine whether a pair of records (one each from birth and CHARS data) referred to the same individual. For example, a pair of records was a high likelihood match if both records agreed on hospital, newborn date of birth, sex, first two letters of last name, first two letters of first name, and zip code. Similarly, other combinations of fields were used to determine whether two records matched. In all, for the newborn-CHARS linkage process there were 36 sets of criteria applied, each being less stringent than the one applied in the previous step. The record pairs were assigned a score indicating the likelihood that they were a match. Lower quality matches were reviewed manually to determine accuracy as were instances where a birth record matched multiple CHARS records. UW BERD documentation indicates that a fairly large number of record pairs underwent manual review.

### CHINCHILLA

The current method replaces the hierarchical criteria with a machine learning process that employs random forests to identify links. As with the UW BERD method, our aim was to find links to CHARS records for both the newborn and the mother. However, unlike the former method, we did not restrict CHARS records by diagnosis codes.

For the newborn-CHARS linkage, we included all CHARS records for patients who were 0 to 18 years old at the time of hospital discharge. The result is the potential for a one-to-many match between a newborn and her CHARS records. In addition to linking the newborn to the CHARS record for the birth event, we sought to link the baby to subsequent hospital visits (if any) until the age of 18 years. As a result of not limiting CHARS records by birth-related diagnosis codes we might see, for example, a newborn linked to the CHARS record for the birth event (with birth-related diagnosis codes), and a second CHARS record for a health issue unrelated to birth outcomes. Linking the newborn to CHARS records beyond the birth event also allows us to examine health issues related to the gestational period and congenital issues.

In the mother-CHARS linkage, we restricted CHARS records to patients whose gender was not male (i.e. female, unknown, or blank) and who were between the ages of 10 and 60 years at the time of discharge. As with newborns, we wanted to link the birth mother to the CHARS record for the birth event as well as any subsequent hospital visits that may have occurred.

## 2014 Comparison

To examine whether the new method was able to match or outperform the performance of UW BERD, we ran the machine learning model on 2014 births, the last year that UW BERD file was created. For this analysis we compared linkage yield for newborns only. We compared UW BERD and CHINCHILLA at various levels. First, we modified the CHINCHILLA process to match the UW BERD process by restricting links to record pairs to those having:

- (1) the same birth-related diagnosis codes used in the UW BERD file, OR
- (2) an interval between the infant’s date of birth and the date of hospital admission of 1 day or less.

Secondly, we compared UW BERD and CHINCHILLA without either of the above restrictions.

For each comparison we did not exclude birth records based on the type of birth facility. Birthing centers, midwiferies, home births, and military hospital births are among the birthplace types that do not report to CHARS thus removing the possibility of linking these births to a birth-related CHARS record. These facility types accounted for about 7% of births in 2014 (n = 6,206). Documentation for UW BERD does not indicate that these birth records were excluded from the denominator when calculating their linkage yield. To keep our accuracy calculations as similar to UW BERD, we used all births as the denominator, but included an additional calculation using the restricted denominator including only CHARS-reporting facilities.

Both methods yielded similar results. UW BERD linked 89% of newborns to their own CHARS records for the birth event as did CHINCHILLA. If other (non-birth event) CHARS records are included, CHINCHILLA captured about 90% of newborns.

	<b>2014 Links</b>	<b>% Total Births (Denominator = 88,450)</b>	<b>% Births from CHARS reporting Facility (Denominator = 82,244)</b>
<b>UW BERD</b>	78,988	89.3%	96.0%
<b>CHINCHILLA Birth Events</b>	79,108	89.4%	96.2%
<b>CHINCHILLA Unrestricted</b>	80,381	90.9%	97.7%

## Limitations

We trained the random forest model used to identify birth-CHARS links using several comparisons of corresponding fields in the birth and CHARS records. In more recent years, some of the most important comparisons, in terms of distinguishing links from non-links, have been the similarity in patients’ residential street address fields. CHARS did not collect patient residential street addresses in 2014,

making a key piece of information unavailable to the model. We hypothesize even greater linkage rates if address information were recorded.

Although the UW BERD method defined birth events in CHARS using ICD-9-CM only codes V30-V39, there are other newborn related codes that may be helpful in identifying birth-related CHARS events when used in conjunction with dates of birth and hospital admission. Congenital anomalies (ICD9-9-CM codes 740 through 759) or conditions originating in the perinatal period (ICD9-9-CM codes 760-779) may identify records where the V codes typically used at birth are inadvertently left off. Similarly, V29 (neonatal observation and evaluation of newborns) or V6405 (vaccination not carried out because of caregiver refusal) may indicate a birth-event related CHARS record. Aside from diagnosis codes, the CHARS variable indicating the type of admission may also be helpful in finding links. Admission code '4' is used to indicate that the patient is a newborn.

## Using CHINCHILLA for analysis

As the methods used in identifying links for UW BERD and CHINCHILLA are so different, we discourage the comparison of rates between data sets generated by the two methods.

CHINCHILLA linkage includes links between birth records and CHARS records for non-birth-events. For this reason, we recommend using the diagnosis codes, interval between dates of birth and admission, and admission type if your analysis is restricted to CHARS records for the birth-event only. We also suggest using other ICD-9-CM codes for 2014 and prior (outlined above) or the ICD-10\_CM equivalents for 2016 onward to identify birth events that may have been miscoded.

For birth years in which CHARS records used ICD-10-CM diagnosis codes a birth event is indicated by codes in the range Z30 to Z38. In addition to these, we recommend looking for codes indicating perinatal conditions (P00 to P96) in conjunction with a short interval between admission and birth dates. Again, the admission type code (code 4) can also be used in conjunction with these secondary diagnosis codes to find the few links that may have been miscoded.

## Appendix A: CHINCHILLA Model Accuracy

To evaluate model accuracy, we selected for manual review between 400 and 500 record pairs that were predicted as links or non-links by the model. These record pairs were edge cases with a probability of being a link on either side of the 0.5 threshold. We also reviewed record pairs assigned extremely high or extremely low probabilities of being a link. For each reviewed pair a label was assigned indicating whether the reviewer agreed with the model or not. The model performed very well in comparison to the reviewer's assigned labels with the following accuracy measures obtained using a 2X2 table.

Measure	
Accuracy	0.995
False detection rate	0.0008
Sensitivity	0.98
Specificity	0.99
Cohen's Kappa	0.99



DOH 422-296 April 2026

To request this document in another format, call 1-800-525-0127. Deaf or hard of hearing customers, please call 711 (Washington Relay) or email [doh.information@doh.wa.gov](mailto:doh.information@doh.wa.gov).