Data Quality Assessment – License Roster: Licensed Practical Nurses in Colorado

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  29 variables in total, including the name, address, city/county, state, zip code, and degree of a licensee, and their license number, type, nurse compact designation, first issue date, last renewed date, expiration date, status description, as well as public action case tracking number, program action levied, effective and complete dates of the action.

- What are the measures of NDC attainment like?
  License. Per BLS definition, a license is a credential awarded by a government agency and conveys a legal authority to work in an occupation.

- Are there any indicators related to education attainment that are unique to this dataset?
  License number, first issue/last renewed/expiration dates, status description, and public/program action records are unique to this administrative dataset.

- Are there indicators of other phenomena that could be of sociological significance?
  No.

- What is the purpose of the dataset, and how closely does that purpose align with the following use cases? (Evaluate as relevant or not relevant.)
  The dataset is intended for license lookup and verification.
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Relevant. Licensee count can be easily obtained from this dataset and used to calculate the attainment rate of a (group of) license within Colorado.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Somewhat relevant. The dataset can be linked to labor market outcome records of the licensees as their names and addresses are given, but such linkage can be challenging as no common unique ID of the licensees (e.g., SSN) is available.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials
     Somewhat relevant. The dataset contains no race/gender information but can be linked to datasets with such information.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Somewhat relevant. The dataset can be linked to labor market outcome records of the licensees.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. The dataset can be used to study the impact of licensure on the specified occupation.
  f. *Other examples we might add?*

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  The dataset attempts to cover all licensed practical nurses in Colorado.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  As of June 16, 2022, the dataset has 47,895 records. Since licensure is a government activity and the dataset comes from an administrative database, we believe this number itself represents the best estimate of the population, given there is no significant withholding of license records in the dataset.
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  Data is collected through internal system updates, and updates are made by licensees and applicants through online services and by individual board/program actions. Collection instruments include internet update, paper form update and manual data entry.
- Do cases that we believe should exist in the microdata actually exist in the data?
  We do not see any sign of missing cases in our assessment.
- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
  Formatted name, license type and license number have no missing value. For other key variables, missing rate is minimal except for degrees. Address line 2 and county have high missing rates probably because relevant information is not applicable to these variables. See below for details.

Variable	Missing Count	Missing Pct	Variable	Missing Count	Missing Pct
Formatted name	0	0	License type	2	0
Address line 1	57	0.12	License number	2	0
Address line 2	45,529	95.06	Nurse compact designation	2	0
City	19	0.04	License first issue date	22	0.05
County	35	0.07	License Last Renewed Date	23	0.05
State	25,818	53.91	License Expiration Date	22	0.05
Zip code	29	0.06	License Status Description	2	0
Zip code +4	29	0.06	Degree(s)	14,914	31.14

3. Granularity

- How granular (i.e., how many different categories exist, if not continuous) is data for key variables of interest (attainment, field of study, income)? What about for different levels of aggregation researchers might consider, such as geography, age, and race?
  - Location: 2-line address, city, state, 9-digit zip code.
  - Degrees: specific degree name.
  - Nurse compact designation: Single state/Multi-state

- License status: Active/Active – Restricted/Active - With Conditions/Expired/Inactive/Revoked/Summary Suspension/Surrendered/Suspended/Voluntary Surrender/Volunteer.
- Program action: specific program action type.

- Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under "relevance"?
  Yes.

4. Timeliness

- How often is the dataset updated?
  Updated in real time.
- Is data collected continuously or in waves? If in waves, what is the duration of those waves? Are some variables/records more frequently updated than others?
  Continuously. Since the data is not centrally updated, records can be updated at different frequencies.
- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  The dataset is updated in real time.

5. Integrity

- What are the risks to the integrity of this dataset?
  Since licensees and applicants can update certain variables by themselves, they may misreport some information (e.g., degree) in their favor.
- How are data outliers handled? (May be available from published documentation if not metadata.)
  No relevant information in technical documentation. Outlier is not a significant issue in this dataset as there is no continuous variable.
- Were there changes to the dataset that may have resulted from political influence? If so, do those changes threaten the overall quality of the data?
  None that we are able to identify.

6. Accessibility

- How do researchers access this dataset?
  The dataset can be downloaded here in Excel/CSV/text formats.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Three variables (sub-category of license, licensee specialty, and licensee title) have no observations in this dataset. We do not know whether relevant information is withheld or non-applicable.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Downloading and using the dataset is free.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  No. License number does not serve as a unique identifier. If a licensee has multiple public disciplinary actions, there will be multiple rows generated for the licensee in order to list each case attached to the license record.
- Is it common? (e.g., a SSN might be higher value than an address, though even name/address might be sufficient to match in some cases).
  No UID.
- Do occupation and industry coding schemes correspond to commonly used frameworks such as O*Net and NAICS? If not, are they well documented in metadata?
  No commonly used occupation/industry codes included.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  We were unable to determine whether there were significant changes over time in the metadata.

- What is the construction of the dataset like? Is the microdata organized in "waves"? Can multiple observations for the same unit of analysis (i.e., person) over time be easily linked together?
  The dataset is cross-sectional. Theoretically, records for the same licensee over time can be linked together through name/location/license number, but historical datasets are not available on Colorado.gov.
- How far back do administrative records from this dataset go?
  We assume that administrative records date back to the establishment of licensure for nurses in Colorado – probably several decades – though it is almost certain that the format of the data has changed over the years.