Optimizing Data Quality and Use

Evaluation Guide and Worksheet

Electronic health record (EHR)-based algorithms have the potential to transform public health surveillance. EHRs can be utilized to automatically extract, analyze, organize and communicate chronic disease data to public health agencies for use in decision-making. While EHR-based surveillance shows great potential, the use of raw EHR data must come with an understanding of the accuracy and quality of the data.

This tool can help you develop a structured and logical way to describe the data quality used to generate metrics for chronic disease surveillance, addressing both data validity and reliability. The data to be evaluated will include both the source EHR data and the derivative public health indicators data derived from the source.

Evaluating EHR Data for Validity and Reliability

The validity of data refers to whether or not the data is measuring what is intended to be measured. It is not the same as reliability: reliability is the extent to which a measurement gives results that are consistent. These concepts do not imply that the data is error-free. Errors found should be within a tolerable range so that the associated risks are not significant enough to cause doubt in finding a conclusion or recommendation based on the data.

There are several considerations to take into account when evaluating the validity and reliability of the source EHR and derivative data for measuring chronic disease prevalence and control:

- Missing Data – Each important data element for determining prevalence and control should be evaluated for the percentage of missing data elements. The percentage of missing data should be compared to an acceptable rate as determined by the project team.
- Structured vs. Unstructured Data – The format of important data elements should be reviewed to determine if those data are stored and transmitted in a structured or unstructured/free-text format. Unstructured data carries a significantly higher risk for data entry error and may require transformation before it becomes usable.
- Data Transformation – When data elements are transformed from data providers to meet certain formats or to consolidate multiple data sources, the possibility for error increases. A systematic review of source to transformed data is needed to determine an error rate and if the rate falls within an acceptable range.
- Duplicate Records – A common issue when aggregating EHR data is de-duplication of patient records. When data from disparate EHR systems in similar geographic locations are aggregated, the possibility of duplicate records increases. A review of the procedures to de-duplicate records is needed.

The *Understanding Clinical Data and Workflow* section offers tools to document data requirements and discrepancies.

Develop and Validate Algorithm for Evaluating Prevalence and Control

To develop an algorithm, you need to establish criteria for the condition of interest based on your EHR data elements.

- What are the clinical or diagnostic criteria?

- Are there any associated prescriptions?
- Which codes are associated with the condition?

The algorithms should be tested against an extract of the EHR data. The sensitivity and positive predictive value of the algorithm should be assessed and the algorithms optimized based on the results. The results of the algorithms should be compared to known local and national rates to determine which outcomes met what was expected.

The following examples illustrate diabetes and hypertension algorithms:

A study example for detecting and distinguishing diabetes type-1 and type-2 included an extract of all diagnosis codes, laboratory test results, and medication prescriptions. The algorithm included any of the following criteria at any time:

1. Hemoglobin $A_{1c} \geq$ 6.5 percent
2. Fasting glucose ≥ 126 mg/dL
3. Prescription for insulin outside of pregnancy
4. ICD-9 code 250.xx on two or more occasions
5. Prescription for one or more of the following medications: glyburide, gliclazide, glipizide, glimepiride pioglitazone, rosiglitazone repaglinide, nateglinide, meglitinide sitagliptin exenatide, pramlintide

The study then determined the sensitivity and positive predictive value of the algorithm and applied the algorithm to prospective data. Utilizing multiple criteria—including prescriptions—allowed the algorithm to identify a broader range of patients.

A hypertension study additionally hypothesized that significantly more cases would be identified through a measure based on the guidelines for diagnosis of hypertension presented in the Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (two or more most recent blood pressure readings greater than or equal to 140/90 mm Hg among those without any diagnosis of essential hypertension). This measure identified those patients who were at risk for undiagnosed hypertension.

The algorithm for the study utilized the following criteria:

6. Identify unduplicated patients with a diagnosis of essential hypertension based on ICD-9-CM codes (using the diagnosis and demographic portions of the data)
7. Identify unduplicated patients with a diagnosis of essential hypertension based on free-text entries (using the diagnosis and demographic portions of the data)
8. Identify unduplicated patients whose last two or more blood pressure readings were greater than or equal to 140/90 mm Hg and who did not have a documented diagnosis of essential hypertension in either ICD-9-CM code or free-text format (using the diagnosis, demographic, and vital signs portions of the data)

Table 1: Data elements needed to identify cases

Determine exactly what data elements your system needs to identify cases for the condition of interest. Include data needed to support specific analyses and generate any desired reports. Document them in the worksheet as well as the standard code system and value set for each data element, when such a standardized vocabulary set exists. Consider developing this data model with your data trading partners (see *Understanding Clinical Data and Workflows* for information on structured vs. unstructured data).

Concept Name	Concept Definition	Code System	Value Set	Explanation of Code	Required