

# Transforming Clinical Data

.....More Accurately than Humans

**Piano's Automapper**

Authored by Dr Guy Tsafnat

A series of approximately 10-12 parallel, wavy purple lines that originate from the bottom left corner and curve upwards and to the right, creating a sense of motion and flow across the lower half of the page.

# Abstract

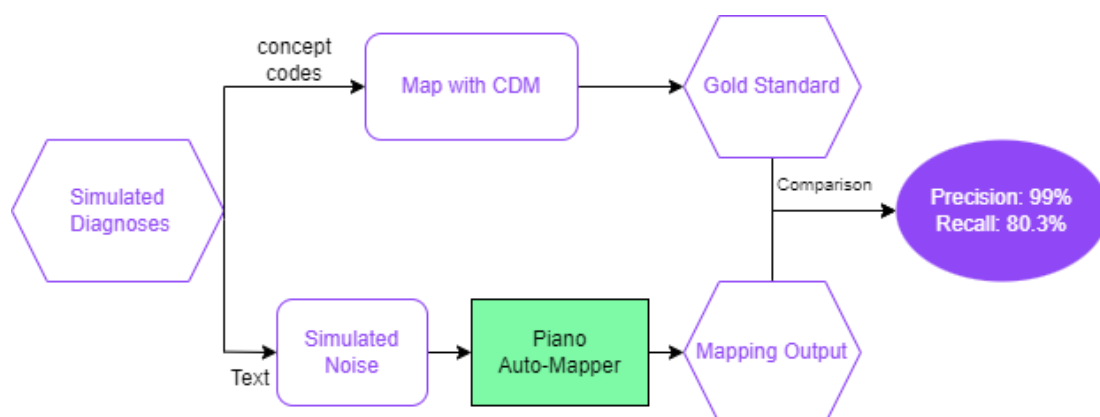
Evidentli's flagship technology, Piano, streamlines and accelerates evidence creation and includes tools that transform data from any source to the Observational Healthcare Data Sciences and Informatics' Common data Model (OHDSI CDM). Together, the combination of AI, automation and clinical user interface design significantly accelerates the transformation from healthcare data to actionable evidence.

However, although the use of AI is faster and economical, it almost inevitably leads to loss of precision compared to human experts. In some domains it is acceptable to trade off accuracy or speed. Healthcare research is not tolerant of low precision as it can lead to devastating consequences for patient outcomes and healthcare costs.

To evaluate Piano's AI in automating the mapping of short-form text to medical concepts from a controlled vocabulary, we conducted an experiment using simulated patient data with simulated errors, and measured the precision of mappings made by the AI.

The Piano Auto-Mapper AI achieved precision of 99%; that is, 99% of the mappings it made were correct. Human experts make twice as many mistakes. More than 80% of the diagnoses were mapped automatically in the first iteration of the algorithm.

## Visual Abstract



# Introduction

Evidentli's flagship research automation platform Piano is an end-to-end solution that uses a number of algorithms and a streamlined user interface design to transform healthcare data into evidence. This includes the transformation of data from any source to the Observational Medical Outcomes Partnership (OMOP) CDM.

A typical data ingestion workflow is called an ETL (extract, transform, load). Extraction of health data from source systems and loading health data to relational databases are tasks that can easily be adapted from other (i.e. non-health) domains. In contrast, transformation of healthcare data requires specific domain knowledge and can lead to erroneous research which can, in turn, harm patients, and waste precious resources. For these reasons, transformation of healthcare data is often done manually, by highly trained individuals, at high costs.

Piano's data ingestion workflows include specific tools to accelerate the transformation of healthcare data into the CDM. These tools streamline the work of human operators performing the task. A major component of the Piano toolset is the Auto-Mapper, an ensemble algorithm that uses machine learning (ML) and natural language processing (NLP) to map data in short-text form into concepts in a standard vocabulary, as well as to map concepts from one vocabulary to another. The mapping of medical concepts is arguably the most time-consuming and error-prone task in healthcare data transformation. Real-world data is notoriously noisy as a result of lack of adherence to a standard as well as differences in medical jargons in addition to common errors such as typos. A recent assessment of data quality loss showed about 2% loss in the manual standardisation of real-world data (Liaw et al. 2020).

# Standardising Clinical Data with Piano

Evidentli's Piano includes workflows for transforming data extracted from a source into the OHDSI OMOP CDM. A data source is typically an electronic medical record system (EMR), patient registry, patient administration system or other system that maintains a patient-level clinical database. The Auto-Mapper is the part of the workflow that uses algorithms to automate the standardisation of medical concepts. Medical concepts include diagnoses, procedure names, test names, anatomical terms, medications and more.

The Auto-Mapper algorithm can take one or two possible inputs. For coded data sources, it takes the source vocabulary's name and the concept codes to be transformed into the standard vocabulary. For uncoded data, the algorithm takes text concepts (see Figure 1). In the case of imperfectly coded data, it takes both the codes and the text descriptions. In all cases the algorithm produces a mapping from the source concept to the standard vocabulary such as SNOMED-CT.

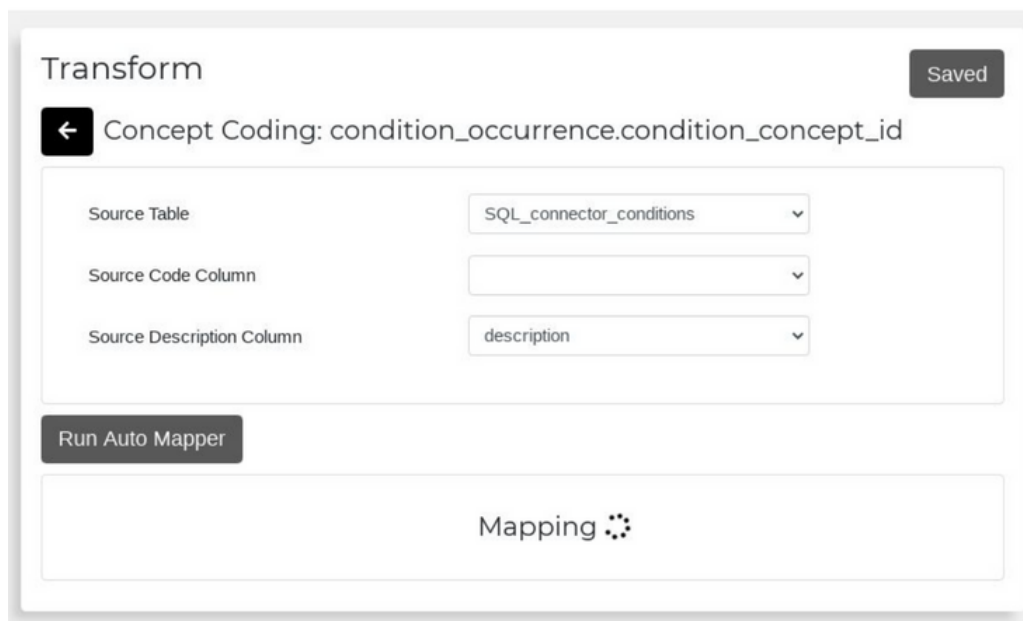
The screenshot shows a web interface titled "Transform" in the top left corner. In the top right corner, there is a dark grey button labeled "Saved". Below the title, there is a navigation bar with a left arrow icon and the text "Concept Coding: condition\_occurrence.condition\_concept\_id". The main area contains three input fields: "Source Table" with a dropdown menu showing "SQL\_connector\_conditions", "Source Code Column" with an empty dropdown menu, and "Source Description Column" with a dropdown menu showing "description". Below these fields is a dark grey button labeled "Run Auto Mapper". At the bottom of the interface, there is a large white box with the word "Mapping" and a circular arrow icon, indicating the process is running.

Figure 1: a screenshot of Piano's concept transformation tool showing the Auto-Mapper running after the user selected the source text field (description) only. In this example, as in this experiment, the source code column is left blank.

The first time the Auto-Mapper is run for a target field, it creates an initial classification model from input data. This step may take several hours, however subsequent iterations of the Auto-Mapper use active learning to incrementally improve this model based on user-provided mappings. At the end of each iteration the Auto-Mapper will present, for each concept found in the source, the concept's mapping if it exists, and whether user input is still required. User input is required if the Auto-Mapper did not find an accurate mapping or if it found more than one (typically two or three) mappings. In the latter case the mappings are presented to the user who can eliminate one or more of the suggested mappings, manually search for a different mapping or accept all mappings (we note that in some cases, one-to-many mappings are appropriate).

# Evaluation

## Data

Seventy five thousand patients' records were generated using Synthea (Walonoski et al. 2018). Synthea was set for the Seattle, Washington locale (as per the Synthea tutorial) with default parameters. The simulation generated 125,757 ICD-9-CM diagnosis codes.

Synthea produces codes and descriptions taken from the ICD-9-CM vocabulary (ICD9CM). The gold standard was created by mapping the Synthea-generated codes using the OHDSI's concept mapping tables. The ICD-9-CM codes were left out of the training and test data so they were not seen by the Auto-Mapper.

Noise is commonly found in real-world datasets. It was simulated using random errors that were introduced to the diagnosis text. Each diagnosis had up to 10 mutations applied. Each mutation was applied to a random character as one of:

- a deletion of the character (P=10%);
- insertion of a random letter right before it (P=10%);
- reversal of its case from lower to upper or vice-versa (P=10%);
- replacing it with a random letter (P=10%); or
- making no further changes (P=60%).

## Mapping Task

Arguably, the most time-consuming task in transforming clinical datasets is transforming short text fields that describe medical concepts such as diagnoses, procedures and allergies. Using a controlled vocabulary greatly simplifies research conducted on medical data. In typical settings, a proficient coder familiar with the controlled vocabulary and often using a vocabulary browser, interprets the phrase and matches it with the closest term from the controlled vocabulary. In this evaluation we simulated coding concepts from the ICD9CM-based simulated dataset described above into the SNOMED-Controlled Terminology standard.

Piano's Auto-Mapper tool for coding text phrases and for translating between vocabularies is an iterative tool that uses machine learning to improve mappings based on corrections made by a human operator until a perfect mapping is achieved. For the purposes of this evaluation we restricted the Auto-Mapper to only the first iteration, before user input is entered.

# Results

The Auto-Mapper can produce a single mapping from a source mapping (Figure 2 a). In this case the concept is labeled as auto-mapped and has a green background. A user can override the mapping. In other cases, the Auto-Mapper may find multiple potential mappings for the concept (Figure 2 b). If so, the concept is mapped as "input needed", the options are listed and the background is red. The user can delete one or all of the suggested mappings and/or provide a manual mapping. Whenever the Auto-Mapper doesn't find a mapping for a concept (Figure 2 c), the concept is also labeled as "input needed", the background is red but there are no suggestions listed. Manually mapped concepts (not shown) are labeled as "Manually mapped" with a green background and are not overridden by the Auto-Mapper.

a				
10,679	Open wound of face, unspecified site, without mention of complication	47126008	Open wound of face without complication	Auto-mapped

b				
5,053	Acute bronchitis and bronchiolitis	5505005	Acute bronchiolitis	Input needed
		111273006	Acute respiratory disease	

c				
1	Diabetes with other specified manifestations	Click here to edit		Input needed

Figure 2. Screen captures of mapping results showing the number of times the concept appeared in the source dataset, the original text description, the mapping or mappings and the status.

The mapping was considered to be correct (i.e. True Positive; TP) if the Auto-Mapper produced a single correct mapping, or if the correct mapping was one of the suggested mappings. The mapping was considered incorrect (i.e. False Positive; FP) if the Auto-Mapper produced a single mapping that was incorrect, or if none of the mappings suggested were correct. The mapping was considered to have failed (i.e. False Negative; FN) if the Auto-Mapper produced no mapping.

The precision achieved by the Auto-Mapper after a single iteration was 98.98%. The recall achieved by the Auto-Mapper was 80.3%. Of those, the Auto-Mapper returned a single mapping in 85.7% of the cases. In cases that the Auto-Mapper returned multiple options for the user to choose from, the correct option was among them in 100% of the cases

$$\text{Precision} = \frac{TP}{TP + FP} = 98.98\%$$

$$\text{Recall} = \frac{TP}{TP + FN} = 80.3\%$$

# Conclusions

The Auto-Mapper provided as part of Piano is a fast and accurate tool to transform patient-level clinical data to the Common Data Model. By comparison with earlier work, the Auto-Mapper only loses approximately half the data quality that human mappers lose. A single iteration of the tool can accelerate the coding of text fields such as diagnoses by a factor of 5. In combination with other elements of Piano and with multiple iterations of the Auto-Mapper, Piano is likely to be able to accelerate data transformation and standardisation by a much greater factor.

# References

(ICD9CM) International Classification of Diseases, Ninth Revision, Clinical Modification, <https://www.cdc.gov/nchs/icd/icd9cm.htm>, last accessed June 18, 2020.

(Liaw et al. 2020) Liaw S-T, Borelli A, Guo G-N, Jonnagaddala J. Data for impact: Does ETL affect their quality? OHDSI Showcase; 2020 May, <https://www.ohdsi.org/2020-eu-symposium-showcase-10/> last accessed June 18, 2020

(Walonoski et al. 2018) Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*. 2018 Mar 1;25(3):230-8.