



# AI powered data mapping automation

Guy Tsafnat, PhD<sup>1,2</sup>

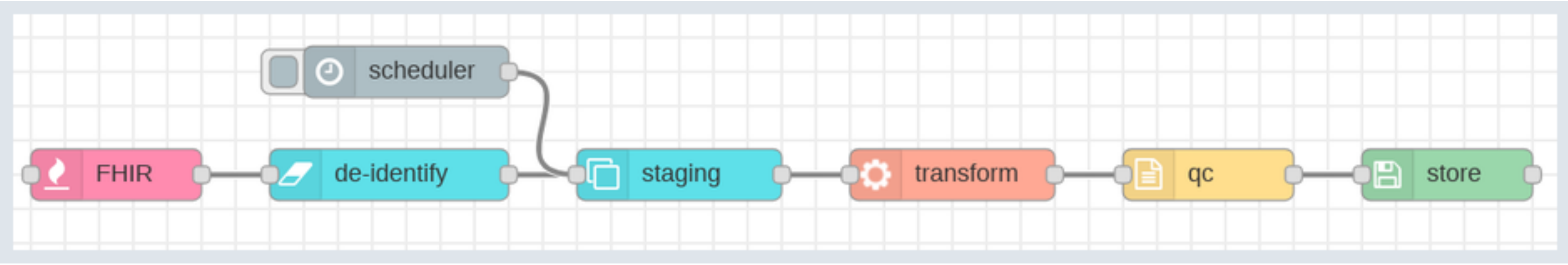
<sup>1</sup> Founder and Chief Science Officer, Evidentli Pty Ltd

<sup>2</sup> Adjunct Fellow, Center for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

## Background

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 expedite high-quality clinical research. This includes the transformation of data from any source to the Common Data Model (CDM) developed and maintained by the international volunteer organization Observational Healthcare Data Sciences and Informatics.

Piano’s data ingress 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).



## Methods

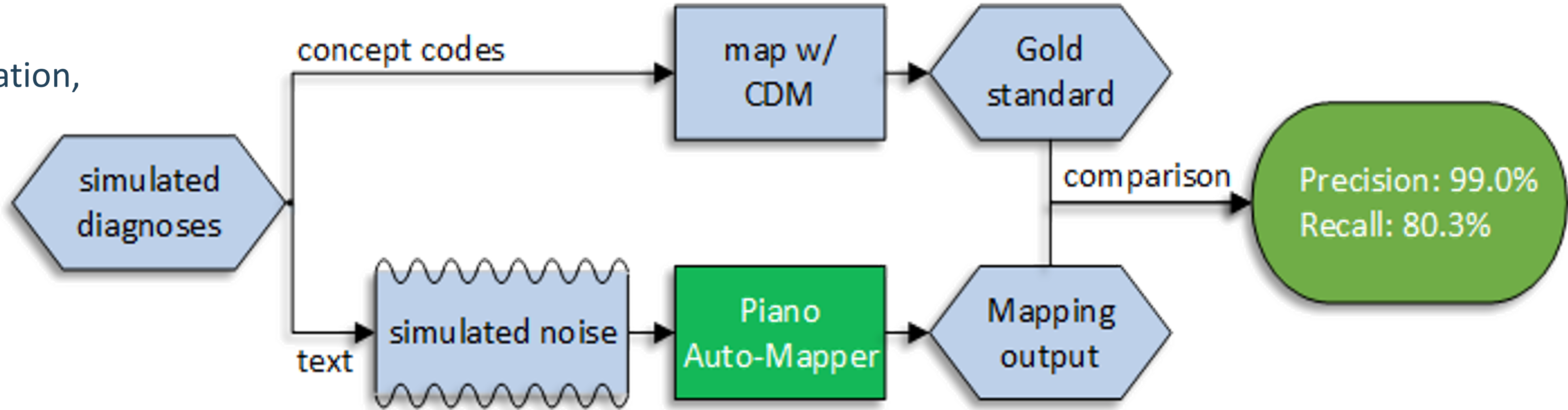
Seventy five thousand patients’ records were generated using Synthea (Walonoski et al. 2018). Synthea was set for the Seattle, Washington locale (as per the Synthea tutorial) with default parameters. The simulation generated 125,757 ICD-9-CM diagnosis codes.

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 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%).

The simulated data and gold standard used in this study are available from Evidentli.com or by request to [info@evidentli.com](mailto:info@evidentli.com).




## Results

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. For examples:

a) a single mapping made by the algorithm,

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

b) multiple mappings are given to the user to choose from, and

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

Click to edit

c) no mapping found by the algorithm, supported by inline search, can manually map the concept.

1	Diabketes with other srpecifiedmanifestations	<a href="#">Click here to edit</a>	Input needed
---	---	------------------------------------	--------------

The mapping was considered to be 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 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 False Negative (FN) if the Auto-Mapper produced no mapping.

Precision  $P=TP/(TP+FP)$  achieved by the Auto-Mapper after a single iteration was **P=98.98%**.

Recall  $R=TP/(TP+FN)$  achieved by the Auto-Mapper was **R=80.3%**.

The Auto-Mapper returned a single mapping in 85.7% of the recalled cases. The correct option was among the suggestions whenever the Auto-Mapper returned multiple options.

## 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 about 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.

(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.