Paper ML13

Assessment of Machine Learning Methods in Coding of Concomitant Medications in Clinical Trials

Sanna Herrgard, Novo Nordisk A/S, Soeborg, Denmark
Carla Gil, Novo Nordisk A/S, Soeborg, Denmark
Ingrid Holst, Novo Nordisk A/S, Soeborg, Denmark
John Holmer Nielsen, Novo Nordisk A/S, Soeborg, Denmark
Mattis Flyvholm Ranthe, Novo Nordisk A/S, Soeborg, Denmark*
Ajser Serif, Novo Nordisk A/S, Soeborg, Denmark
Damon Fahimi, Uppsala Monitoring Centre, Uppsala, Sweden
Jesper Kjaer, Novo Nordisk A/S, Soeborg, Denmark

ABSTRACT

In this talk we will discuss automation of coding of concomitant medications in clinical trials using machine learning. We have assessed performance of WHODrug Koda, developed by Uppsala Monitoring Centre (UMC), with data from 12 different studies across different therapeutic areas. We conclude that Koda would enable Novo Nordisk to increase automation level in the concomitant medication coding process from the current 62% to 79% when coding to Drug Code on the Trade Name level. Out of the 21% of the data that did not have a high certainty Drug Code prediction from Koda, 15% did have a single or multiple lower certainty Drug Code predictions which may provide additional value in the coding process. We observe that high certainty Drug Codes and ATC (Anatomical Therapeutic Chemical) codes predicted by Koda have a high conformance with Novo Nordisk coding with 96% and 94% agreement respectively. A detailed analysis of the differences in coding reveals that Koda codes frequently more in detail through its use of Indication. The results suggest that machine learning supports automation of concomitant medication coding.

INTRODUCTION

Application of machine learning (ML) methods is making its way to clinical development. One possible application is automation of concomitant medication coding. Coding of concomitant medications is a well-defined area, where a lot of suitable training data exists within sponsors. Many patients in clinical trials tend to have similar concomitant medications, depending on therapeutic area. Finally, correctness of predictions from ML-based methods can be easily verified.

In this study we have performed an assessment of the WHODrug Koda service (https://www.who-umc.org/whodrug/whodrug-portfolio/whodrug-koda/) from Uppsala Monitoring Centre (UMC). WHODrug Koda is an automated coding service for coding of concomitant medications against the WHODrug Global dictionary. Koda performs both Drug Name coding and ATC selection and expects user to provide Verbatim Term, Indication and Route of administration as input (Figure 1). For the Drug Name coding, Koda identifies key components from the raw data, applies a series of spelling checks, algorithms, coding rules meeting industry accepted coding conventions and, to some extent, also ML. For the ATC selection, a supervised ML model, based on Natural Language Processing (NLP) and logistic regression is used. The ML model applied for Koda is trained using drug data within VigiBase – the WHO global database of individual case safety reports (ICSRs), developed and maintained by UMC. Koda is continuously retrained for each WHODrug Global release.

Verbatim Term	Indication	Predicted Drug Name (Koda)	Predicted Drug Code (Koda)	Predicted ATC code (Koda)
Aspirin 81 mg tablet	to help prevent heart attack and/or stroke.	ASPIRIN	00002701004	B01AC, Platelet aggregation inhibitors excl. heparin
Aspirin	Headache	ASPIRIN	00002701004	N02BA, Salicylic acid and derivatives

Assigned by Investigator

Prediction from Koda

Figure 1. An example of results from Koda. Aspirin is an example of a concomitant medication where multiple ATC codes exist depending on Indication.

Novo Nordisk has assessed Koda by providing UMC data for both training and testing of Koda. Performance of Koda was assessed both before and after training Koda with Novo Nordisk data. Results were analyzed by comparing predictions from Koda to coding results from Novo Nordisk's current coding process which is presented in Figure 2. Stage 1 of the coding process consists of matching Verbatim Term assigned by the Investigator to the WHODrug B3 Global dictionary. This stage is performed automatically using a synonym list in the Thesaurus Management System (TMS). If no match is found in the synonym list, coding is performed manually. Stage 2 consists of selecting an ATC code manually for those concomitant medications that have multiple ATC codes (i.e. ATC selection). Performing stage 2 may require a medical background, and final judgements typically are conditioned on information about Indication which is collected through Concomitant Medication Form as well as related medical history and adverse events. An ML-based automation method should be able to support automation of both stages of coding, i.e. coding to Drug Code and selecting a single ATC code.

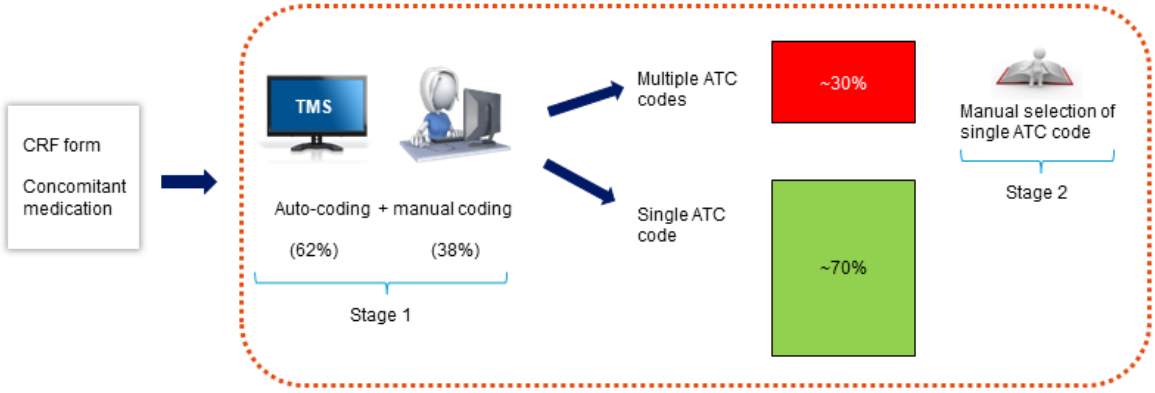

Figure 2. An overview of Novo Nordisk's current coding process.

CHARACTERIZATION OF DATA SETS

Figure 3 summarizes the data sets that were used to train and test Koda. Concomitant medication data from a total of 12 clinical trials over different therapeutic areas (type 1 and type 2 diabetes, obesity and haemophilia) were used. The data sets contained a total of 93,448 data points. All of the data sets had been through the current coding process at Novo Nordisk, where the Verbatim Term was matched to a Trade Name in the WHODrug B3 Global dictionary. Additionally, all but one trial (Trial #1 in Fig. 3) had also been through the ATC selection process.

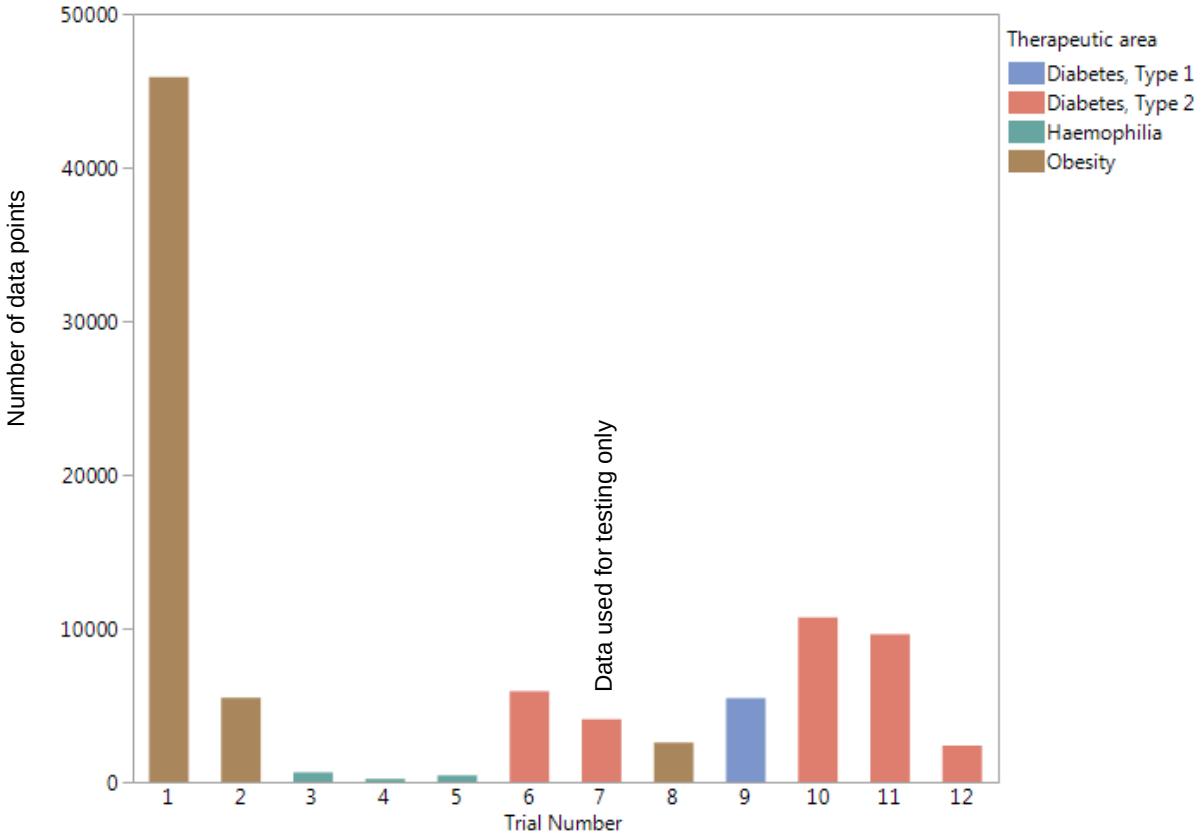

Figure 3. Data sets used in the training and testing of Koda.

Data from 11 trials were used to train and test the performance of Koda. For each data set, 70% of data were randomly selected for training and 30% were used for testing. Additionally, data from one trial (Trial #7) was used for testing only.

All training and test data sets contained Verbatim Term and Indication. Route was not available.

Training process is outlined in Figure 4. Novo Nordisk data was used to train Koda by adding the Novo Nordisk data to the predefined and already existing training data set. Terminology Specialists at UMC also manually provided more selective training material for the ML model based on examples where the original predictions by Koda on the training data sets were incorrect.

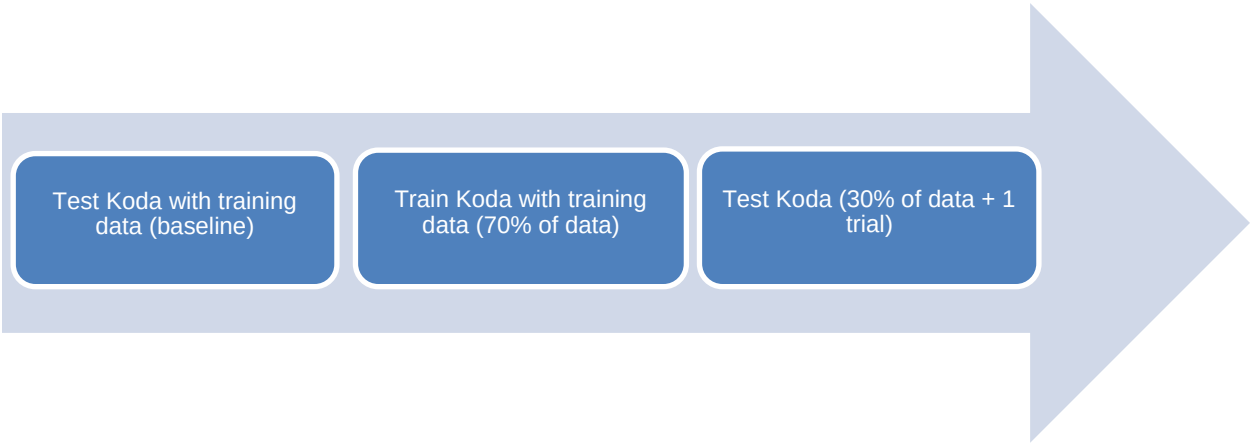

Figure 4. Training process of Koda with Novo Nordisk data

RESULTS FOR TEST DATA SETS

DRUG CODE PREDICTION EFFICIENCY

Drug Code prediction efficiency is displayed in Figure 5. Figure 5 demonstrates that Koda predicted a single high certainty Drug Code for 79% of the test data. Predictions from Koda are accompanied with a confidence level. For the sake of optimal usability, Koda translates the calculated confidence levels into high and lower certainty Drug Code predictions. For an additional 15% of the test data, Koda predicted a single or multiple lower certainty Drug Codes and for 6% of the test set there was no Drug Code prediction from Koda. These 21% of the data would require manual intervention from Novo Nordisk in either confirming a lower certainty prediction, selecting a single Drug Code from multiple possible, or doing coding manually. As compared to Novo Nordisk auto-coding rate of 62% this is a significant improvement.

Drug Code prediction efficiency was also separately analyzed for Trial #7 where none of the data from the trial was used for training. Koda predicted a single high certainty Drug Code for 88% of the data set and for an additional 8% of the data set Koda predicted a single or multiple lower certainty Drug Codes. This illustrates the fact that patients within different therapeutic areas have different concomitant medications which can lead into varying prediction efficiencies.

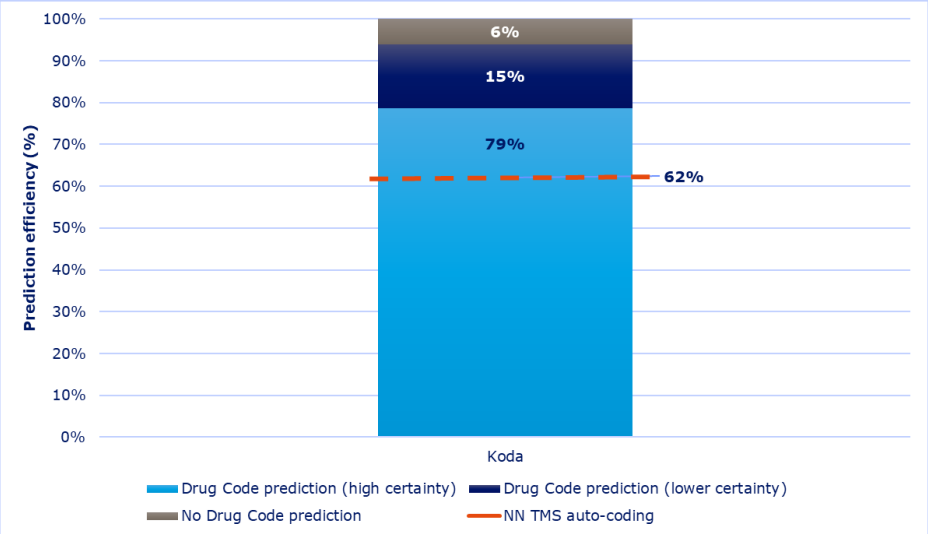

Figure 5. Drug Code prediction efficiency by Koda.

It is impossible for any automated solution to select a single Drug Code in all cases. Figure 6 contains an example for Claritin for nasal congestion. Koda predicts three different options with different active ingredients, Drug Codes and ATC codes. In this case Coder needs to select the correct entry based e.g. on where the patient is located or based on active ingredients.

Verbatim Term	Indication	Lower certainty Drug Name (Koda)	Drug Code (Koda)	Ingredients (Koda)	High certainty ATC code (Koda)	Lower certainty ATC code (Koda)
Claritin	nasal congestion	CLARITIN /00984601/	00984601641	Clarithromycin		D06AX
		CLARITIN /00917501/	00917501011	Loratadine	R06AX	
		CLARITIN /00413701/	00413701085	Gliclazide	A10BB	

Figure 6. Example of predicted coding from Koda for Claritin for nasal congestion, with three possible options for Claritin with three different active ingredients, Drug codes and ATC codes.

DRUG CODE PREDICTION CONFORMANCE

Conformance of predicted Drug Codes was analyzed by comparing Drug Codes for all high certainty predictions from Koda, i.e. 79% of the data set, to Drug Codes from Novo Nordisk coding. Figure 7 shows that for 96% of the data set, Koda coding agrees with Novo Nordisk coding.

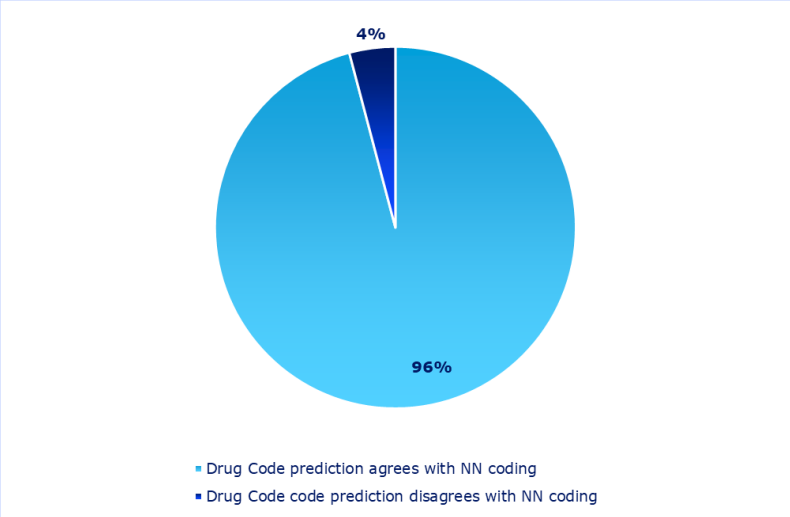

Figure 7. Comparison of Drug codes between Koda coding and Novo Nordisk coding for all high certainty predictions.

In order to understand the differences between Koda coding and Novo Nordisk coding on a deeper level, a sample of 181 concomitant medications with disagreeing Drug Code coding were analyzed in detail. Figure 8 shows that in 46% of the cases, coding from both Koda and Novo Nordisk were equally acceptable; for 44% Koda was able to code more in detail, and in 10% of the cases Novo Nordisk coding was more precise. One of the reasons why Koda codes more in detail is that Koda uses Indication in coding, whereas Novo Nordisk does not use Indication when coding to a Trade Name. Indication is used by Novo Nordisk only with ATC selection.

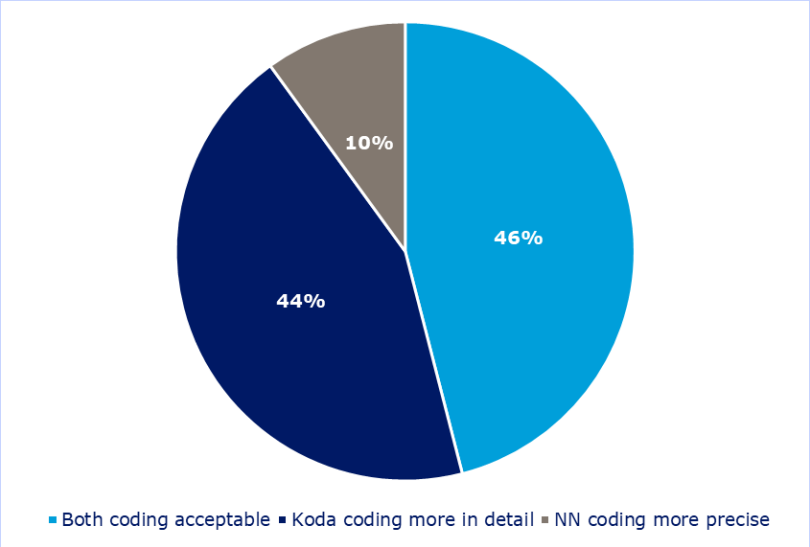

Figure 8. Analysis of 181 concomitant medications with disagreeing Drug Code coding between Koda and Novo Nordisk.

A detailed analysis was performed of the high certainty Drug Code predictions where discrepancies were observed between Koda coding and Novo Nordisk coding. While majority of the differences can be explained by multiple entries of e.g. vitamins in the WHODrug Global dictionary or different salt forms, we also observe that for instance coding of combination drugs (drugs with multiple active ingredients) may need to be confirmed manually in certain cases. This is not due to a deficiency in the performance of Koda, but due to the different ways combination drugs can be collected using Concomitant Medication form. Combination drugs can be collected either via a single entry in the Concomitant Medication form, or each active ingredient separately in conjunction with the combination drug (Figure 9). If active ingredients are collected as separate entries in conjunction with the combination drug (option 2), coding should ideally be to individual active ingredients as opposed to the combination drug. It is to be noted here that different sponsors probably have different ways of collecting combination drugs.

Option 1:
Generic or Trade Name of combination drug

Option 2:
Generic or Trade Name of combination drug [active ingredient 1]
Generic or Trade Name of combination drug [active ingredient 2]

Figure 9. Two different options for collecting combination drugs using Concomitant Medication form.

ATC CODE PREDICTION EFFICIENCY

Koda's efficiency on predicting ATC codes is shown in Figure 10. The prediction efficiency is slightly lower as compared to prediction efficiency of Drug Codes. This is understandable as there can be multiple ATC codes per Drug Code corresponding to different uses of the concomitant medication, and Indication needs to be considered when selecting the correct ATC code. Additionally, it was observed in some cases that if Route of administration would have been available, Koda would have been able to predict the correct ATC code.

We also analyzed ATC code prediction efficiency separately on Trial #7. Similarly to the Drug Code prediction efficiency, Koda was able to predict a single high certainty ATC code for 82% of the concomitant medications in this trial.

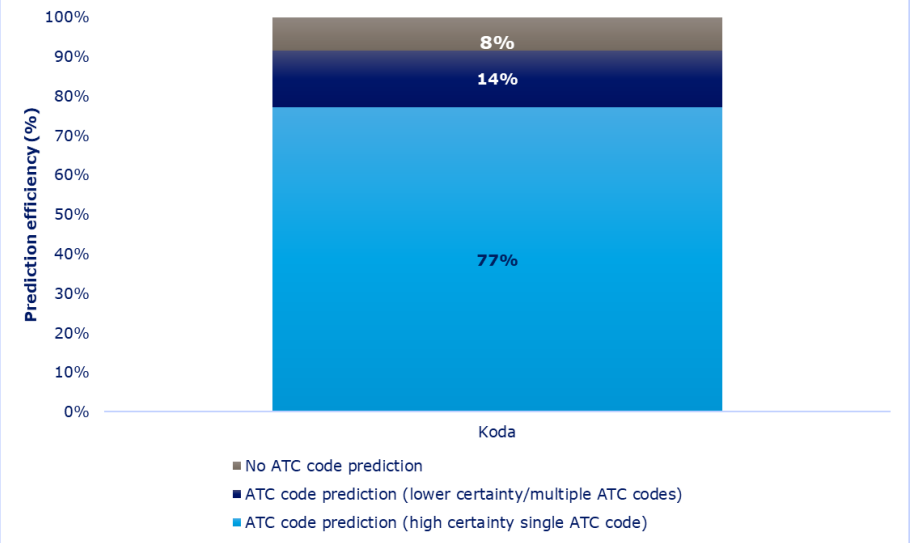

Figure 10. ATC code prediction efficiency by Koda.

ATC CODE PREDICTION CONFORMANCE

Conformance of predicted ATC codes was analyzed by comparing ATC codes for all high certainty predictions from Koda to ATC codes from Novo Nordisk coding. This corresponds to 77% of the test data set from Figure 10. Results from this comparison are presented in Figure 11. Figure 11 demonstrates that ATC codes from Koda and Novo Nordisk coding agreed for 94% of the test data set. For 5% of the data set ATC codes disagreed and for 1% there was no ATC code from Novo Nordisk coding. These were mainly cases where Coder had considered that additional information would have been needed to select a single ATC code.

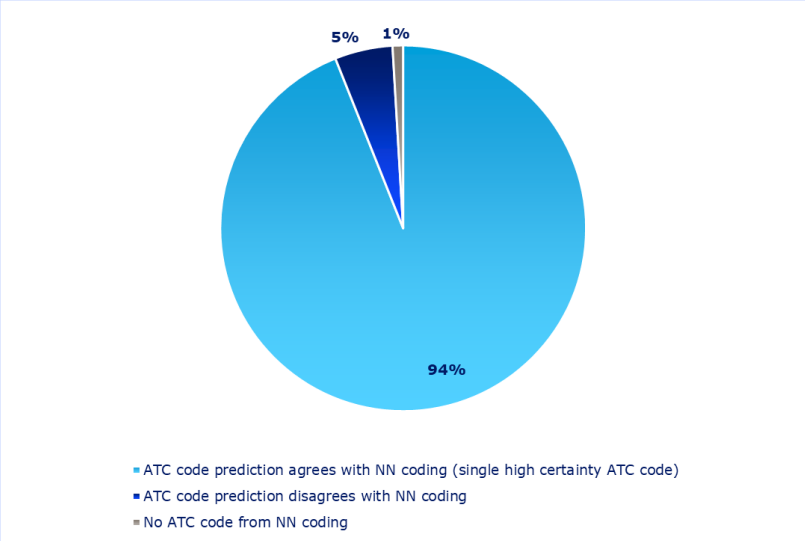

Figure 11. Comparison of ATC codes between Koda coding and Novo Nordisk coding for all high certainty predictions.

CONCLUSION

Results from this assessment demonstrate that it would be feasible for Novo Nordisk to increase automation level in the concomitant medication coding process from the current 62% to 79% when coding to Drug Code on the Trade Name level. Out of the 21% of the data that did not have a high certainty Drug Code prediction from Koda, 15% did have single or multiple lower certainty Drug Code predictions which may provide additional value in the coding process. We observe that high certainty Drug Code and ATC code predictions from Koda have a high conformance with Novo Nordisk coding with 96% and 94% agreement with Novo Nordisk coding. A detailed analysis of the differences reveals that in many cases Koda codes more in detail through use of Indication in coding. Our analysis shows that combination drugs may require manual confirmation, depending on how combination drugs are collected using the Concomitant Medication form. Results from this assessment suggest that machine learning supports automation of concomitant medication coding.

One of our key learnings from this assessment is the importance of understanding your data. It is important to spend time not only understanding the method, but also the results and where the predictions may go wrong. It is through this detailed analysis that makes it feasible for sponsors to take advantage of ML-based methods in clinical trials.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name Sanna Herrgard
Company Novo Nordisk A/S
Address Vandtaarnsvej 114
City / Postcode 2860 Soeborg, Denmark
Work Phone: +45-30798271
Email: sanh@novonordisk.com

*Current address: ALK Abello, Hoersholm, Denmark