Using a Convolutional Neural Network to Predict Remission of Diabetes After Gastric Bypass Surgery: Machine Learning Study From the Scandinavian Obesity Surgery Register

Yang Cao1,2, PhD; Ingmar Näslund3, MD, PhD; Erik Näslund4, MD, PhD; Johan Ottosson3, MD, PhD; Scott Montgomery1,5,6, PhD; Erik Stenberg3, MD, PhD

1Clinical Epidemiology and Biostatistics, School of Medical Sciences, Örebro University, Örebro, Sweden
2Unit of Integrative Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
3Department of Surgery, Faculty of Medicine and Health, Örebro University, Örebro, Sweden
4Division of Surgery, Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Stockholm, Sweden
5Clinical Epidemiology Division, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
6Department of Epidemiology and Public Health, University College London, London, United Kingdom

Corresponding Author:
Yang Cao, PhD
Clinical Epidemiology and Biostatistics
School of Medical Sciences
Örebro University
X-Huset, Södra Grev Rosengatan 1
Örebro, 70182
Sweden
Phone: 46 196 026 236
Fax: 46 196 026 236
Email: yang.cao@oru.se

Abstract

Background: Prediction of diabetes remission is an important topic in the evaluation of patients with type 2 diabetes (T2D) before bariatric surgery. Several high-quality predictive indices are available, but artificial intelligence algorithms offer the potential for higher predictive capability.

Objective: This study aimed to construct and validate an artificial intelligence prediction model for diabetes remission after Roux-en-Y gastric bypass surgery.

Methods: Patients who underwent surgery from 2007 to 2017 were included in the study, with collection of individual data from the Scandinavian Obesity Surgery Registry (SOReg), the Swedish National Patients Register, the Swedish Prescribed Drugs Register, and Statistics Sweden. A 7-layer convolution neural network (CNN) model was developed using 80% (6446/8057) of patients randomly selected from SOReg and 20% (1611/8057) of patients for external testing. The predictive capability of the CNN model and currently used scores (DiaRem, Ad-DiaRem, DiaBetter, and individualized metabolic surgery) were compared.

Results: In total, 8057 patients with T2D were included in the study. At 2 years after surgery, 77.09% achieved pharmacological remission (n=6211), while 63.07% (4004/6348) achieved complete remission. The CNN model showed high accuracy for cessation of antidiabetic drugs and complete remission of T2D after gastric bypass surgery. The area under the receiver operating characteristic curve (AUC) for the CNN model for pharmacological remission was 0.85 (95% CI 0.83-0.86) during validation and 0.83 for the final test, which was 9%-12% better than the traditional predictive indices. The AUC for complete remission was 0.83 (95% CI 0.81-0.85) during validation and 0.82 for the final test, which was 9%-11% better than the traditional predictive indices.

Conclusions: The CNN method had better predictive capability compared to traditional indices for diabetes remission. However, further validation is needed in other countries to evaluate its external generalizability.

(JMIR Med Inform 2021;9(8):e25612) doi: 10.2196/25612

KEYWORDS
forecasting; clinical decision rules; remission induction; type 2 diabetes mellitus; gastric bypass; morbid obesity
Introduction

Bariatric surgery is an efficient and safe treatment for patients with morbid obesity and type 2 diabetes (T2D) [1,2]. In obese patients who also have T2D, more than three-fourths of patients show remission after gastric bypass surgery [3,4]. Although remission rates may differ across different surgical procedures, high remission rates have been reported for Roux-en-Y gastric bypass [1,3]. Despite many patients experiencing remission of diabetes, duration and severity of disease, along with age, have been presented as factors associated with reduced chance of achieving remission [1,5]. Prediction of diabetes remission can be helpful in clinical preoperative consultation and decision-making, and several indices have been constructed for this purpose. Scores like DiaRem [6], Ad-DiaRem [7], DiaBetter [8], and the individualized metabolic surgery (IMS) score [9], as well as the age, body mass index, C-peptide level, and duration of T2D (ABCD) score [10] have been used for predicting diabetes remission after bariatric surgery. Many of the models based on the scores have high predictive capability and may already provide clinical guidance [11]. These tools might be helpful for personalized management of morbidly obese individuals with diabetes when considering bariatric surgery in routine care, ultimately contributing to precision medicine [12]. However, the performance of the scores in various studies is not consistent [7]. Previous prediction models were either limited by small sample sizes or were not validated using external data that were not seen by the models during model construction. Therefore, both the performance and validity of the models or scores need to be further evaluated and improved using a larger bariatric surgery database. In recent years, there have been a number of attempts to use artificial intelligence (AI) algorithms, including support vector machine [13], decision tree [14], random forest [15], and deep learning algorithms, such as artificial neural networks [16,17], to incorporate preoperative predictors in predicting outcomes of bariatric surgery. Compared with the traditional statistical regression models, AI algorithms have shown great promise in the field of bariatric surgery [18,19]. However, to our knowledge, none have thus far reached clinical practice.

The aim of this study was to construct a prediction model for T2D remission using a deep learning AI algorithm (ie, convolutional neural network [CNN]) and to compare its predictive capability with that of 4 widely used predictive scores.

Methods

Study Participants

The study used the data from the Scandinavian Obesity Surgery Register (SOREg), a validated, national quality register covering virtually all bariatric and metabolic surgical procedures in Sweden [20]. By using the unique Swedish personal identification number, we linked SOReg to the Swedish National Patient Register, the Swedish National Death Register, the Swedish Prescribed Drug Register, and Statistics Sweden to obtain information on inpatient and outpatient hospital visits, mortality, dispensed drugs, and individual socioeconomic data. The inclusion criteria for patients registered in the SOReg were included those operated on with a primary Roux-en-Y gastric bypass procedure between 2007 and 2015 and those diagnosed with T2D preoperatively, as defined by the American Diabetes Association (ie, fasting plasma glucose ≥ 126 mg/L [7.0 mmol/L], hemoglobin A1c [HbA1c] ≥ 48 mmol/mol [6.5%], or pharmacological treatment for diabetes) [21].

Outcome and Predictor Variables

The main outcome measure was complete remission of diabetes 2 years after surgery, defined as being without diabetes medication within a time frame of +/- 6 months; that is, 18-30 months postoperatively with normal HbA1c value <42 mmol/mol (6.0%) in accordance with the definitions of the American Diabetes Association [22]. Due to loss of information of HbA1c at follow-up, analyses of a secondary outcome, complete remission, defined as discontinuance of pharmacological treatment from 18-30 months, was performed.

The predictor variables were patients’ demographic and socioeconomic information including age, sex, education level (primary, secondary, higher education <3 years, and high education ≥3 years), and region of residence characteristics (large city, medium city or town, and small town or rural area); preoperative BMI, HbA1c, and treatment information including insulin treatment, metformin use, other noninsulin pharmacological treatment, and number of antidiabetic drugs; and preoperative comorbidities including sleep apnea, hypertension, dyslipidemia, depression, and cardiovascular comorbidity.

Descriptive Analysis

Continuous variables are presented as mean and SD, and ordered and nominal variables are presented as median and interquartile range (IQR) and count and percentage, respectively. For comparison between 2 groups, the t test and Mann-Whitney test were used for continuous and ordered variables, respectively, while the Pearson chi-square test was used for categorical variables. A 2-tailed P value <.05 was considered to be statistically significant.

Multiple Imputation for Missing Values

Missing values were assumed missing at random and imputed using a random forest algorithm, which has the desirable properties of being able to handle mixed types of missing data, being adaptive to interactions and nonlinearity, and having the potential to scale to big data settings [23]. To allow for the uncertainty of the imputation, 100 imputed data sets were generated in the current study.

Data Normalization

Because the range of values of variables varies widely (such as for age and BMI) in some machine learning (ML) algorithms, objective function will not work properly [24]. Therefore, the continuous and ordered variables were normalized to have a mean of 0 and a standardization of 1, and the multivariate nominal variables (education and residence) were converted into several binary variables before they were entered into the ML models [25].
**Predictive Model**

In the current study, we used a 7-layer CNN model with two 1D convolution layers (with 100 filters for each), two 1D max pooling layers, one flatten layer, and two dense layers (with 1000 computation units) [26,27]. The rectified linear unit activation function was used for the two 1D convolution layers and the first dense layers, and the sigmoid activation function was used for the last dense layer. The binary cross-entropy loss function and the adaptive moment estimation (Adam) optimizer were used when compiling the model [28].

**Model Training, Validation, and Test**

The whole data set was randomly split into 2 parts: a training data set with 80% (6446/8057) of the patients and a test data set with 20% (1611/8057) of the patients. During the model training stage, the training data set was further divided into 2 data sets: one data set with 64% (5156/8057) of the patients to train the CNN model and another with 16.01% (1290/8057) of the patients to validate the model. Finally, the model was tested using the test data set that was never seen by the CNN model. The CNN model was trained, validated, and tested with the 100 imputations (Figure 1).

**Indices of Predictive Ability**

Predictive ability of the CNN model was evaluated using the following indices: area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, and the Youden J [29]. The terminology and derivations of the values have been previously presented in detail [18]. The sensitivity and specificity presented in this study are the values on the ROC curve where the Youden J achieves the maximum value. The acceptable, excellent, and outstanding predictive models were defined as those with an area under the ROC curve (AUC) greater than 0.7, 0.8, and 0.9, respectively [30,31]. The average and the 95% CI of the indices were calculated based on 100 imputations.

**Comparison Between the CNN Model and DiaRem, Ad-DiaRem, DiaBetter, and IMS**

We also evaluated the predictive capability of the currently used indices, DiaRem, Ad-DiaRem, DiaBetter, and IMS, and compared them with the CNN model. The DiaRem score is calculated using insulin use, age, HbA\textsubscript{1c} value, and type of antidiabetic drugs [32]. The Ad-DiaRem score is a modification of the DiaRem score, calculated using insulin use, age, HbA\textsubscript{1c} value, number of antidiabetic drugs, duration of diabetes, and number of antidiabetic drugs [13]. The DiaBetter is calculated using HbA\textsubscript{1c} type of antidiabetic drugs, and duration of diabetes [8]. The IMS score is calculated using the number of preoperative diabetes medications, insulin use, duration of diabetes, and HbA\textsubscript{1c} level [9].

The points on the nonparametric ROC curve of DiaRem, Ad-DiaRem, DiaBetter, and IMS were generated using each value as a classification cutoff point and computing the corresponding sensitivity and one minus specificity. These points were then connected by straight lines, and the AUC was computed using the trapezoidal rule [33].

The same training and testing procedure used for the CNN model was also applied for the 4 scores.

**Software and Hardware**

The descriptive analysis and evaluation for DiaRem, Ad-DiaRem, DiaBetter, and IMS were conducted in Stata 16.1
The CNN model was achieved in Python 3.6 (Python Software Foundation) using the Keras 2.4.0 and Scikit-learn 0.23 packages. All the computation was operated on a computer with 64-bit Windows 7 Enterprise operating system (Service Pack 1, Microsoft Corporation), an Intel Core TM i5-4210U 2.40-GHz CPU, and 16.0 GB installed random access memory.

Ethics
The study was approved by the regional ethics committee in Stockholm (reference #2013/535-31/5, #2014/1639-32, and #2017/857-32). The study was conducted according to the guidelines of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement [34].

Results

Patient Characteristics
In total, 8112 patients met the inclusion criteria; after exclusion of 55 patients who died within the first 2 years after surgery, 8057 patients remained in the analysis. Information on pharmaceutical usage before and after surgery was available for all patients. A postoperative weight was registered for 7268 patients at 1 year after surgery (90.21%), and 4996 patients at 2 years after surgery (62.01%). A postoperative glycosylated HbA1c test result was available for 6989 patients (86.74%). Baseline characteristics of the included patients are shown in Table 1. Statistically significant differences were found for almost all the predictor variables between the remission patients and nonremission patients, except for depression and education (Table 1), which implies the potential for using the predictor variables to predict outcome. Preoperative HbA1c values were missing for about one-seventh of the patients, indicating the need for imputation since the predictive capability otherwise would be significantly reduced and biased by excluding a considerable proportion of the data with missing values. Patients with a missing HbA1c value were more often males of marginally higher age and longer duration of disease, and small differences were also seen in terms of pharmacological treatment, education, and residence (Supplementary Table S1, Multimedia Appendix 1). After multiple imputation, similar distributions of HbA1c values were seen (Supplementary Figure S1, Multimedia Appendix 1).
Table 1. Characteristics of study participants with further stratification on remission of diabetes (N=8057)a.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Overall (n=8057)</th>
<th>Nonremission (n=1846)</th>
<th>Remission (n=6211)</th>
<th>P valueb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>47.7 (10.1)</td>
<td>51.7 (8.7)</td>
<td>46.6 (10.2)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.001</td>
</tr>
<tr>
<td>Women</td>
<td>4970 (61.68)</td>
<td>1079 (58.45)</td>
<td>3891 (62.65)</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>3087 (38.32)</td>
<td>767 (41.55)</td>
<td>2320 (37.35)</td>
<td></td>
</tr>
<tr>
<td><strong>BMI (kg/m²), mean (SD)</strong></td>
<td>42.22 (5.74)</td>
<td>41.16 (5.44)</td>
<td>42.53 (5.80)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Hemoglobin A1c (mmol/mol) mean, (SD)</strong></td>
<td>59.0 (17.3)</td>
<td>67.4 (17.5)</td>
<td>56.7 (16.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diabetes duration (years), median (IQR)</td>
<td>2.0 (0.0-6.0)</td>
<td>6.0 (3.0-10.0)</td>
<td>1.0 (0.0-4.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Number of drugs, median (IQR)</td>
<td>1.0 (1.0-2.0)</td>
<td>2.0 (1.0-2.0)</td>
<td>1.0 (0.0-2.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Insulin, n (%)</td>
<td>2313 (28.71)</td>
<td>1184 (64.14)</td>
<td>1129 (18.18)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Metformin, n (%)</td>
<td>5610 (69.63)</td>
<td>1618 (87.65)</td>
<td>3992 (64.27)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Other noninsulin treatment, n (%)</td>
<td>1912 (23.73)</td>
<td>745 (40.36)</td>
<td>1167 (18.79)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sleep apnea, n (%)</td>
<td>1529 (18.98)</td>
<td>383 (20.75)</td>
<td>1146 (18.45)</td>
<td>.03</td>
</tr>
<tr>
<td>Hypertension, n (%)</td>
<td>4546 (56.42)</td>
<td>1287 (69.72)</td>
<td>3259 (52.47)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Cardiovascular comorbidity, n (%)</td>
<td>917 (11.38)</td>
<td>305 (16.52)</td>
<td>612 (9.85)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Dyslipidemia, n (%)</td>
<td>2527 (31.36)</td>
<td>864 (46.80)</td>
<td>1663 (26.78)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Depression, n (%)</td>
<td>1297 (16.10)</td>
<td>311 (16.85)</td>
<td>986 (15.88)</td>
<td>.34</td>
</tr>
<tr>
<td><strong>Education, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.40</td>
</tr>
<tr>
<td>Elementary education</td>
<td>1606 (19.93)</td>
<td>392 (21.24)</td>
<td>1214 (19.55)</td>
<td></td>
</tr>
<tr>
<td>Secondary education</td>
<td>4762 (59.10)</td>
<td>1091 (59.10)</td>
<td>3671 (59.10)</td>
<td></td>
</tr>
<tr>
<td>Higher education &lt;3 years</td>
<td>838 (10.40)</td>
<td>179 (9.70)</td>
<td>659 (10.61)</td>
<td></td>
</tr>
<tr>
<td>Higher education &gt;3 years</td>
<td>796 (9.88)</td>
<td>173 (9.35)</td>
<td>623 (10.03)</td>
<td></td>
</tr>
<tr>
<td><strong>Residence, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.001</td>
</tr>
<tr>
<td>Large city</td>
<td>2734 (33.93)</td>
<td>687 (37.22)</td>
<td>2047 (32.96)</td>
<td></td>
</tr>
<tr>
<td>Medium-sized town</td>
<td>3061 (37.99)</td>
<td>671 (36.35)</td>
<td>2390 (38.48)</td>
<td></td>
</tr>
<tr>
<td>Small town or rural area</td>
<td>2231 (27.69)</td>
<td>487 (26.38)</td>
<td>1744 (28.08)</td>
<td></td>
</tr>
<tr>
<td>DiaRem, median (IQR)</td>
<td>6.0 (3.0-13.0)</td>
<td>16.0 (8.0-18.0)</td>
<td>5.0 (3.0-8.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Ad-DiaRem, median (IQR)</td>
<td>7.0 (5.0-11.0)</td>
<td>12.0 (9.0-15.0)</td>
<td>7.00 (4.0-9.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DiaBetter, median (IQR)</td>
<td>3.0 (1.0-6.0)</td>
<td>7.0 (5.0-8.0)</td>
<td>3.0 (1.0-4.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>IMS, median (IQR)</td>
<td>39.8 (16.0-75.2)</td>
<td>87.2 (59.9-107.2)</td>
<td>28.6 (16.0-57.8)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

aIncluding all the baseline variables used in the study.
bP value comparing remission vs nonremission.
cIMS: individualized metabolic surgery.

Surgical Outcome

The mean BMI loss at 1 year after surgery was 12.2 kg/m² (SD 4.0 kg/m²), with an excess BMI loss (100 × [initial BMI – postoperative BMI]/[initial BMI – 25]) of 74.0% (SD 22.5%), and a total weight loss (100 × weight loss/preoperative weight%) of 28.7% (SD 7.6%). Mean BMI loss at 2 years after surgery was 12.0 kg/m² (SD 4.53 kg/m²) with an excess BMI loss of 73.3% (SD 24.4%) and a total weight loss of 28.4% (SD 8.9%). At 2 years after surgery, 77.09% (6211/8057) of the patients were able to discontinue the pharmacological treatment of T2D, while complete T2D remission was seen in 63.07% (n=4004) of the 6348 patients who had been evaluated for complete remission.

Predictive Capability of the CNN Model, DiaRem, Ad-DiaRem, DiaBetter, and IMS

The predictive capability of the CNN model for the major outcome (remission) is shown in Figure 2 and Table 2. In both the training and validation, the CNN model presented good predictive ability, with an AUC of 0.86 (95% CI 0.85-0.87) and 0.85 (95% CI 0.83-0.86), respectively (Table 2).
Figure 2. Receiver operating characteristic (ROC) curves of the convolutional neural network model in one of the 100 trainings and validations (left; because the 2 areas under the ROC curves are almost totally overlapping, the blended red and blue colors appear purple), and tests (right).

The DiaRem, Ad-DiaRem, DiaBetter, and IMS also showed good predictive capability in the training with an AUC >0.8 (Figure 3 left and Table 2) but only acceptable predictive ability in the validation (Table 2), with an AUC of 0.73 (95% CI 0.71-0.75), 0.72 (95% CI 0.69-0.74), 0.75 (95% CI 0.72-0.78), and 0.76 (95% CI 0.73-0.79), respectively. In general, the predictive capability of the CNN model was 16.4%, 18.1%, 13.3%, and 11.8% higher than that of DiaRem, Ad-DiaRem, DiaBetter, and IMS, in terms of AUC, respectively. In the tests, the AUC for the predictive ability of the CNN (AUC=0.83; 95% CI 0.82-0.85) model was 10.6%, 12.2%, 12.2%, and 9.2% higher than that of DiaRem (AUC=0.75; 95% CI 0.73-0.76), Ad-DiaRem (AUC=0.74; 95% CI 0.71-0.77), DiaBetter (AUC=0.74; 95% CI 0.72-0.76), and IMS (AUC=0.76; 95% CI 0.73-0.78), respectively (Figure 2 right and Figure 3 right).
Table 2. Predictive capability of the CNN model and diabetes indices for the major outcome.

<table>
<thead>
<tr>
<th>Models by index</th>
<th>Value (95% CI)</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.86 (0.85-0.87)</td>
<td>0.85 (0.83-0.86)</td>
<td></td>
</tr>
<tr>
<td>DiaRem</td>
<td>0.81 (0.79-0.82)</td>
<td>0.73 (0.71-0.75)</td>
<td></td>
</tr>
<tr>
<td>Ad-DiaRem</td>
<td>0.82 (0.81-0.83)</td>
<td>0.72 (0.69-0.74)</td>
<td></td>
</tr>
<tr>
<td>DiaBetter</td>
<td>0.82 (0.81-0.83)</td>
<td>0.75 (0.72-0.78)</td>
<td></td>
</tr>
<tr>
<td>IMS</td>
<td>0.84 (0.83-0.85)</td>
<td>0.76 (0.73-0.79)</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.78 (0.74-0.83)</td>
<td>0.78 (0.72-0.85)</td>
<td></td>
</tr>
<tr>
<td>DiaRem</td>
<td>0.76 (0.80-0.73)</td>
<td>0.81 (0.78-0.85)</td>
<td></td>
</tr>
<tr>
<td>Ad-DiaRem</td>
<td>0.70 (0.68-0.71)</td>
<td>0.75 (0.70-0.79)</td>
<td></td>
</tr>
<tr>
<td>DiaBetter</td>
<td>0.76 (0.74-0.78)</td>
<td>0.76 (0.71-0.80)</td>
<td></td>
</tr>
<tr>
<td>IMS</td>
<td>0.77 (0.72-0.82)</td>
<td>0.77 (0.72-0.81)</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.77 (0.73-0.82)</td>
<td>0.76 (0.70-0.83)</td>
<td></td>
</tr>
<tr>
<td>DiaRem</td>
<td>0.75 (0.71-0.78)</td>
<td>0.65 (0.62-0.67)</td>
<td></td>
</tr>
<tr>
<td>Ad-DiaRem</td>
<td>0.79 (0.78-0.80)</td>
<td>0.69 (0.67-0.72)</td>
<td></td>
</tr>
<tr>
<td>DiaBetter</td>
<td>0.75 (0.74-0.76)</td>
<td>0.75 (0.72-0.78)</td>
<td></td>
</tr>
<tr>
<td>IMS</td>
<td>0.75 (0.70-0.80)</td>
<td>0.75 (0.73-0.77)</td>
<td></td>
</tr>
<tr>
<td>Youden J</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.56 (0.54-0.57)</td>
<td>0.54 (0.50-0.59)</td>
<td></td>
</tr>
<tr>
<td>DiaRem</td>
<td>0.51 (0.50-0.52)</td>
<td>0.46 (0.42-0.50)</td>
<td></td>
</tr>
<tr>
<td>Ad-DiaRem</td>
<td>0.48 (0.47-0.49)</td>
<td>0.44 (0.39-0.49)</td>
<td></td>
</tr>
<tr>
<td>DiaBetter</td>
<td>0.51 (.049-0.54)</td>
<td>0.51 (.045-0.56)</td>
<td></td>
</tr>
<tr>
<td>IMS</td>
<td>0.52 (0.50-0.54)</td>
<td>0.52 (0.47-0.57)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)AUC: area under the receiver operating characteristic curve.  
\(^b\)CNN: convolutional neural network.  
\(^c\)IMS: individualized metabolic surgery.

Figure 3. Receiver operating characteristic curves of diabetes indices in one of the 100 trainings (left), and tests (right). AUC: area under the curve.
For the secondary outcome, complete remission, the CNN model also presented a good predictive capability in both the training and validation, with an AUC of 0.84 (95% CI 0.83-0.85) and 0.83 (95% CI 0.81-0.85), respectively (Supplementary Table S4, Multimedia Appendix 1). Although DiaRem, Ad-DiaRem, DiaBetter, and IMS showed good predictive ability in the training with an AUC ≥0.80, they only showed acceptable predictive ability in the validation with an AUC of 0.72 (95% CI 0.69-0.75), 0.72 (95% CI 0.69-0.74), 0.74 (95% CI 0.72-0.77), and 0.74 (95% CI 0.72-0.76), respectively (Supplementary Table S4, Multimedia Appendix 1). In general, the predictive capability of the CNN model was 15.3%, 15.3%, 12.2%, and 12.2% higher than that of DiaRem, Ad-DiaRem, DiaBetter, and IMS, in terms of AUC, respectively.

In the tests, the AUC for the predictive capability of the CNN model (AUC=0.82; 95% CI 0.81-0.83) was 9.3%, 10.8%, 10.8%, and 9.3% higher than that of DiaRem (AUC=0.75; 95% CI 0.73-0.78), Ad-DiaRem (AUC=0.74; 95% CI 0.73-0.75), DiaBetter (AUC=0.74; 95% CI 0.71-0.76), and IMS (AUC=0.75; 95% CI 0.73-0.77), respectively (Supplementary Figure S2 right and S3 right, Multimedia Appendix 1).

Discussion

Principal Findings

The CNN model evaluated in this study showed high accuracy for cessation of antidiabetic drugs and complete remission of T2D after gastric bypass surgery, providing 9%-12% better predictive indices compared to available scores.

The currently available and widely accepted predictive indices for diabetes remission, including DiaRem, Ad-DiaRem, DiaBetter, and IMS, were assessed in our study and are all simple and easily available to clinicians for clinical decision support. In addition, one other index, the ABCD score [35], also includes c-peptide. This laboratory measure additionally includes information of endogenous insulin production and could thus potentially further enhance the effectiveness of a prediction model. However, the ABCD score has not been shown to have higher predictive capacity compared to other available models, and it is highly possible that other measures of severity of T2D disease, such as duration of disease, HbA1c value, and type and number of drugs, may provide the same or even better measures for a prediction model [11].

The main benefits of the CNN method, in comparison to the scores based on traditional statistical methods, lie in its ability to include a high number of variables and to learn over time. In contrast to available models designed to offer simple entry and calculations of the most important variables, it offers the ability to handle variables in a more complex way, also including variables of smaller impact. Furthermore, the model construction is not limited by the statistical assumptions and distribution of the data, which usually need to be fulfilled in the traditional regression methods. Exposing the AI to a higher quantity of real-world data also has the potential to further improve it with cumulative learning.

Implications

The use of AI or machine learning techniques in medical research and practice is currently an evolving field with great potential. Although the exact role of AI in this setting remains to be established, one potential area where the AI seems to outperform traditional techniques is indeed in the construction of prediction models for outcomes from surgical procedures [36]. Previous studies on the construction of prediction models for perioperative complications have reported discouraging results, mainly as a direct consequence of the complexity and diversity of causes for perioperative complications [18,27,37]. In contrast with safety outcomes, efficacy outcomes (in particular those of highly standardized surgical methods such as gastric bypass) may be more suited for adequate prediction models since the factors influencing long-term effects are less diverse. Remission of diabetes is one such outcome that is largely influenced by a few specific factors, making prediction models more easily available. The results of our study support the promising results from previous studies with smaller sample sizes using sparse support vector machine, decision tree, and artificial neural networks to predict diabetes remission after bariatric surgery [13,14,17].

Although our CNN model did not include postoperative weight loss, a factor known to be associated with higher remission and reduced relapse of diabetes [8], the model included measures of patient-specific characteristics, information on duration and severity of disease, and a few socioeconomic factors that all should be easily available at the time of consultation before surgery. Although it is likely that the model could have reached a higher precision if postoperative results (such as early weight loss or improvement in glucose homeostasis) were included, these measures are not available in the preoperative setting and their inclusion would therefore reduce the clinical usefulness of the model [1,5,38]. Age, duration of diabetes, preoperative HbA1c, and diabetes medications are all known predictive factors [1,5]. In addition, the model identified sex, BMI, metabolic and cardiovascular comorbidities, and place of residence as factors influencing the chance of diabetes remission.

Although the disposition of adiposity and insulin resistance appears to affect men and women differently [39], differences between sexes may be highly influenced by other covarying factors, such as obesity-related comorbidities, BMI, and age [1]. Indeed, when adjustment is made for other factors, the influence of sex on outcome tends to shift [1]. The influence of BMI on remission rates is also controversial [40]. Patients with higher BMI may have a greater degree of insulin resistance and a higher expected total weight loss [41,42], and may thus benefit more from the favorable metabolic effect of bariatric surgery. However, the influence of BMI on remission can be related to several other factors of relevance for both diabetes remission and postoperative weight loss. Whether or not the influence of BMI is strictly weight dependent or not remains to be answered. Although no difference in remission dependent on educational level was seen, place of residence was associated with the chance of achieving diabetes remission. Residents of larger cities may experience higher life stress and represent a more diverse socioeconomic population [43]. Many socioeconomic factors (such as education, income, profession, and ethnicity)
have been reported to influence other efficacy outcomes, such as weight loss, which in turn may contribute to these differences [42].

**Challenges and Limitations**

In contrast to traditional regression models, we observed significant improvement with the continuous training process. When increasing amounts of data in the test data set were seen by the CNN model (or more data in the test data set leaked into the training data set), AUC, specificity, and sensitivity increased gradually and eventually approximated 1 (Supplementary Figure S5, Multimedia Appendix 1). From training with more available data and correlating data with methods such as principal component analysis, the predictive capability of the CNN model could be improved even further, at least in the Swedish context. To generalize the application of the CNN model, a multinational registration consortium of gastric bypass surgery patients would be needed for improved model training and validation. However, the capacity of memory is also a limitation of the CNN because it reduces the model’s flexibility to incorporate the information from external unseen data, which results in overfitting to specific past data or underfitting to the new data and impedes generalization of the model [44]. Teaching neural networks to strategically forget is an important task in ML. This highlights one of the major challenge of ML techniques [45]. To fulfill this task, incorporating long short-term memory units into CNN networks has been attempted to process temporal sequences and reduce model parameters in human face and activity recognition, which has shown consistent superior performance and good generalization [46,47]. Furthermore, the methods of ML are less transparent and more complex than those of traditional regression models, making their exact nature more difficult to scrutinize [44]. In the absence of clear guidelines, we have—to the best of our ability—conducted and reported the study to match the requirements of the TRIPOD statement and suggested modifications [34]. The programming code of the study is available at the repository figshare website [48].

**Conclusions**

Our CNN-based ML model performed well in identifying morbidly obese patients with T2D who might benefit from Roux-en-Y gastric bypass surgery. We also demonstrated the model had better predictive capability compared with the current widely used 4 comprehensive indices for diabetes remission after gastric bypass surgery. Prospectively identifying this subset of patients using data available at the time of preoperative evaluation provides an opportune time window to intervene and prevent or reduce the risk of morbidity and mortality, and may potentially reduce the total cost of care. However, this model should be further validated in future research using external data in other countries before it is incorporated into clinical practice.

**Conflicts of Interest**

JO has received reimbursement for participating in the advisory board of Johnson & Johnson and Vifor PHarma. ES received reimbursement by Johnson & Johnson Medical for a lecture on a topic unrelated to the contents of the present work. All other authors declare no conflicts of interest.

**Multimedia Appendix 1**

Supplementary materials.
[DOCX File, 437 KB-Multimedia Appendix 1]

**References**

Abbreviations

ABCD: age, body mass index, C-peptide level, and duration of type 2 diabetes
AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
HbA1c: hemoglobin A1c
IMS: individualized metabolic surgery
ML: machine learning
SOR: Scandinavian Obesity Surgery Registry
T2D: type 2 diabetes
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 10.11.20; peer-reviewed by J Hjelmesæth, R Krukowski; comments to author 24.11.20; revised version received 07.12.20; accepted 10.07.21; published 19.08.21

Please cite as:
Cao Y, Näslund I, Näslund E, Ottosson J, Montgomery S, Stenberg E. Using a Convolutional Neural Network to Predict Remission of Diabetes After Gastric Bypass Surgery: Machine Learning Study From the Scandinavian Obesity Surgery Register.
JMIR Med Inform 2021;9(8):e25612
URL: https://medinform.jmir.org/2021/8/e25612
doi: 10.2196/25612
PMID: 34420921

©Yang Cao, Ingmar Näslund, Erik Näslund, Johan Ottosson, Scott Montgomery, Erik Stenberg. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.