Classifying Continuous Glucose Monitoring Documents From Electronic Health Records

Yaguang Zheng, PhD[1] (iD), Eduardo Iturrate, MD[2],
Lehan Li, MS[3] (iD), Bei Wu, PhD[1], William R. Small, MD, MBA[2],
Susan Zweig, MD[4], Jason Fletcher, PhD[1], Zhihao Chen, MS[5] (iD),
and Stephen B. Johnson, PhD[6]

Abstract

Background: Clinical use of continuous glucose monitoring (CGM) is increasing storage of CGM-related documents in electronic health records (EHR); however, the standardization of CGM storage is lacking. We aimed to evaluate the sensitivity and specificity of CGM Ambulatory Glucose Profile (AGP) classification criteria.

Methods: We randomly chose 2244 (18.1%) documents from NYU Langone Health. Our document classification algorithm: (1) separated multiple-page documents into a single-page image; (2) rotated all pages into an upright orientation; (3) determined types of devices using optical character recognition; and (4) tested for the presence of particular keywords in the text. Two experts in using CGM for research and clinical practice conducted an independent manual review of 62 (2.8%) reports. We calculated sensitivity (correct classification of CGM AGP report) and specificity (correct classification of non-CGM report) by comparing the classification algorithm against manual review.

Results: Among 2244 documents, 1040 (46.5%) were classified as CGM AGP reports (43.3% FreeStyle Libre and 56.7% Dexcom), 1170 (52.1%) non-CGM reports (eg, progress notes, CGM request forms, or physician letters), and 34 (1.5%) uncertain documents. The agreement for the evaluation of the documents between the two experts was 100% for sensitivity and 98.4% for specificity. When comparing the classification result between the algorithm and manual review, the sensitivity and specificity were 95.0% and 91.7%.

Conclusion: Nearly half of CGM-related documents were AGP reports, which are useful for clinical practice and diabetes research; however, the remaining half are other clinical documents. Future work needs to standardize the storage of CGM-related documents in the EHR.

Keywords

diabetes mellitus, continuous glucose monitoring, natural language processing, electronic health record

[1]Rory Meyers College of Nursing, New York University, New York, NY, USA
[2]Department of Medicine, Grossman School of Medicine, New York University, New York, NY, USA
[3]Center for Data Science, New York University, New York, NY, USA
[4]Division of Endocrinology, Department of Medicine, Grossman School of Medicine, New York University, New York, NY, USA
[5]Department of Biostatistics, School of Global Public Health, New York University, New York, NY, USA
[6]Department of Population Health, Grossman School of Medicine, New York University, New York, NY, USA

Corresponding Author:
Yaguang Zheng, PhD, Assistant Professor, Rory Meyers College of Nursing, New York University, 433 1st Avenue, New York, NY 10010, USA.
Email: yaguang.zheng@nyu.edu

Recent data from the Centers for Disease Control and Prevention have shown an increase in the prevalence of diabetes in the United States of ~14.7%, or approximately 37.1 million people.[1] Direct and indirect costs attributed to diabetes in the United States were $327 billion in 2017.[2] Substantial evidence shows that sustained blood glucose control (HbA$_{1c}$ < 7%) is significantly associated with a decreased incidence of diabetic complications.[3] However, more than 70% of patients have failed to meet recommended HbA$_{1c}$ goals.[4]

Continuous glucose monitoring (CGM) shows promise as a way to improve glycemic control. Numerous studies have shown that using CGM has improved glycemic control and quality of life for individuals with diabetes.[5-9] The use of CGM devices among patients with diabetes has been increasing and grew significantly during the period of time dominated by the COVID-19 pandemic.[10] The CGM data are a powerful source to inform the clinical management of diabetes.[11] The CGM allows both patients and clinicians to visualize glucose levels and patterns in real-time and also allows them to look back at patterns and trends over the weeks and months, which can inform daily decision-making regarding how to balance food, physical activity, and medications.[12,13] To be most useful, CGM data need to be collected, transmitted, presented, stored, and processed in near real-time, as well as be integrated into the clinician's main electronic health record (EHR) workflow, and are contextualized with the rest of the patient's clinical data.[11,14]

The most useful CGM-related data are from the Ambulatory Glucose Profile (AGP).[15] The AGP is a report that summarizes CGM data over multiple days of wear.[16] The international consensus statement has reached an agreement on using the AGP as the default report for CGM.[16] There are several core components of AGP reports, including but not limited to data completeness (eg, percentage of active wearing days), glucose level statistics (eg, time above, in, and blow range), the glucose management indicator (GMI), and glucose variability. The AGP helps clinicians and individuals with diabetes understand, assess, and optimize diabetes management, making it highly valuable in clinical practice and diabetes research.

Presently, there is a lack of standardization for collecting, transmitting, presenting, storing, and processing CGM data in the EHR by health care systems, which may cause less efficient workflows to use CGM data and potentially contribute to increased provider burnout.[8] The research is limited with regard to how CGM data are displayed and consumed in the EHR. The purpose of this study was to (1) classify CGM documents existing in the EHR from a large academic health care system and (2) describe the current status of CGM presentation and storage in clinical practice at a large urban academic medical center. This work is an important step in understanding the management of CGM data in the EHR, which is helpful for the timely delivery of glucose data to providers or automated reports based on CGM data. This work can contribute to using CGM data effectively and efficiently to improve the quality of glycemic control for patients with diabetes.

Methods

The NYU Langone Health (NYULH) uses a single EHR system across five inpatient hospitals and >350 ambulatory locations throughout the greater metropolitan New York City area, including Manhattan, Brooklyn, and Long Island. We queried the EHR for documents uploaded to patients' charts that contained keywords related to CGM use in the file names. The data set of documents obtained by this approach contained CGM AGP reports as well as other uploaded documents related to CGM use. These documents originally were entered into patients' charts using two different workflows. One set of documents consists of scanned documents, and the other set of documents was electronically transmitted documents from CGM applications. Because a manual process was involved, the scanned documents contained variability in alignment on the page, with some documents being upside down, for example. The electronically transmitted documents, on the contrary, were uniform in page alignment across all documents. The study was approved by the institutional review board at NYU LH.

Steps of Document Classification

We retrieved CGM-related documents from the EHR at NYU LH 2012-2022. We randomly chose 2244 (18.1%) out of 12 415 documents and classified which documents were true CGM AGP reports that provide glucose visualization. Our document classification algorithm includes four steps: (1) separate multiple-page document into a single-page image; (2) rotate all pages into an upright orientation; (3) apply optical character recognition to the image; and (4) test for the presence of particular keywords in the text. The classification algorithm (Figure 1) was developed using Python. The document classification pipeline illustrates this algorithm as a series of processing steps (Figure 2).

Step 1. Separate multiple-page document into a single-page image: Each file contains multiple pages of a patient's glucose records. The algorithm uses a pre-trained computer vision model to perform optical character recognition (OCR) and extract text from page images (keras_ocr, built in tensorflow). This package can only handle one image at a time, so the multi-page file is converted into several single-page images.
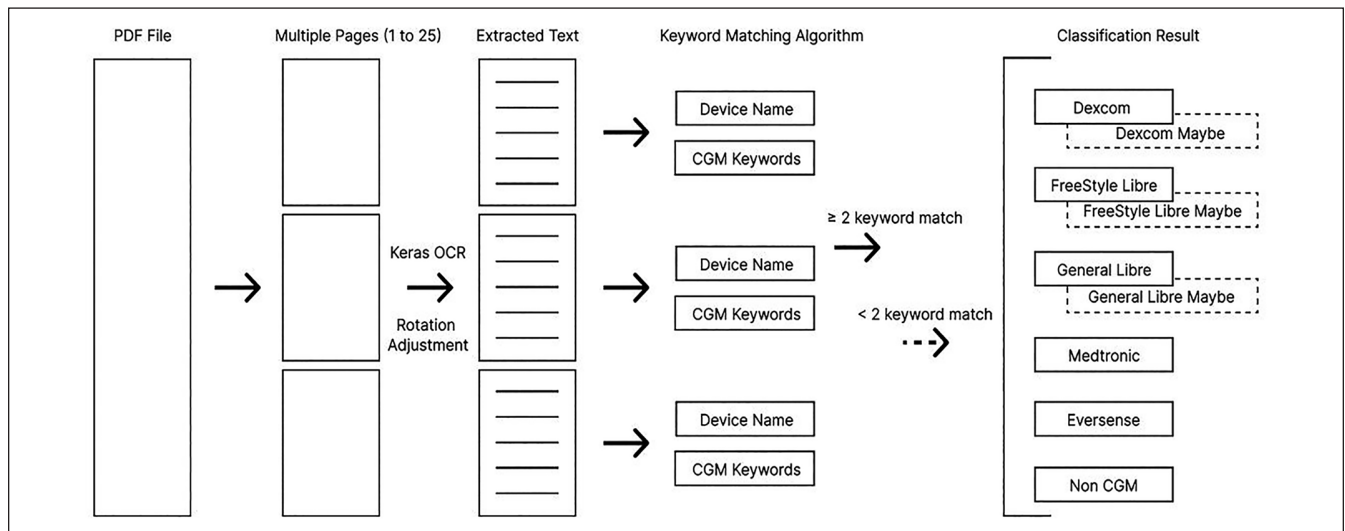
Step 2. Rotate document: The page images are not necessarily in the upright position; some of them are upside down or rotated 90 degrees. We did not consider the case when the image is only slightly rotated because the model is able to recognize the text if the rotation angle is small. A manual review of the data suggests that files with incorrect positioning are only in the format of 90- or 180-degree rotation. Since the text extraction model typically works

Figure 1. Document classification algorithm.

Figure 2. Document classification pipeline.

best when the image is in the upright position, this step detects the position of the image and rotates it to the correct position.[17] The model takes an image as input, extracts the text, and outputs a list of words. If the image is not in the upright position, the extracted text will usually contain only empty strings, one-letter words and two-letter words.

To determine the position of the document, the program takes the extracted text and returns the proportion of words with a length of two or less. We compared extracted text from both upright positions, 90 degrees clockwise, 90 degrees anticlockwise, and 180 degrees clockwise, and we compared the proportion of words with a length of two or less. We determined that if the proportion of very small words is greater than 70%, then the image is considered to be in the incorrect position. This method ascertains whether the image is upright but is unable to determine how to rotate it to obtain the correct image. We therefore rotate the original image 90 degrees clockwise, 90 degrees anticlockwise, and 180 degrees clockwise to produce three different images. We apply a word-length metric to all three rotated images and select the one with the lowest proportion.

Table 1. Key Words Used for the Document Classification Algorithm.

Device type	Keywords for device type	Keywords for CGM
FreeStyle	Contains both - "freestyle" - "libre" or "librer"	- daily log - average glucose - time in target
LibreView	- Doesn't contain "freestyle" - contain "libreview"	- modal day - average tests - glucose history - glucose management indicator - glucose variability - time cgm is active - snapshot - low glucose event - mealtime patterns - weekly summary - device details - daily patterns - daily glucose summary
Dexcom	- contain "dexcom"	- average glucose - time in range - hypoglycemia risk - statistics for this date range - pattern for this date range
Medtronic	- contain "medtronic"	
Eversense	- contain "eversense"	

Step 3. Determine the type of device using optical character recognition: The goal of this step is to determine what type of CGM device was used to generate the document, such as Dexcom G6 or Freestyle Libre, or whether the document is not an AGP CGM report at all. Each document file can contain "useful" pages (ie, containing CGM AGP report) and "useless" pages (ie, not containing CGM AGP report). Sometimes, all the pages inside a file are useful CGM records, while other times, useful pages are mixed with useless pages. The algorithm processes all the pages in the file and classifies the file as a useful CGM document if a single useful image is detected.

The name of the CGM device typically appears in the header or footer of useful pages of the document. However, in some files, the top or the bottom of the text on the page is missing (especially for PNG and TIF files), which are scanned from paper copies or photos taken from paper copies. Different devices include different sections in their reports, such as "time in range" and "hypoglycemia risk." The algorithm uses the presence of keywords that appear in section names as a criterion to classify the type of device.

Step 4. Test for the presence of particular keywords in the text: The code loops through all the single-page images for a file and stops when one image is considered a useful

CGM report or passes through all the images without finding a useful report.

The device type is updated for each image unless the device type is not recognized (when the device type of the current page is missing). Because useful CGM reports are typically contiguous, if no device type is detected, the current page will inherit the device type from the previous page.

According to the device type of the current page, a list of keywords (phrases) is used to match the extracted text (Table 1). Since the extracted text is a list of single words, the code matches the phrases word by word in the list. It is possible that the list of extracted words contains every single word in the keyword phrase, but the words come from different locations and do not form a phrase in the original file. We determined that a given page can be considered useful if the extracted text contains two or more keywords (phrases). When this threshold is reached, the algorithm classifies the document as the corresponding device type and moves on to the next file. If the page is not useful, the algorithm continues to the next page. If all pages have been checked, the algorithm classifies the document as a non-CGM file.

Algorithm Evaluation

Two experts in using CGM for research and/or clinical practice conducted an independent manual review of 62 (2.8%) reports out of 2244. We calculated sensitivity (correct classification of CGM AGP report) and specificity (correct classification of non-CGM report) by comparing the classification algorithm against manual review.

Results

Among 2244 documents, 1040 (46.5%) were classified as CGM AGP reports (43.3% FreeStyle Libre series and 56.7% Dexcom G series), 1170 (52.1%) non-CGM reports (eg, progress notes, CGM request forms, or physician letters), and 34 (1.5%) uncertain documents. The agreement for the evaluation of the documents between the two experts was 100% for sensitivity and 98.4% for specificity. When comparing the classification result between the algorithm and manual review, the sensitivity and specificity were 95.0% and 91.7%, respectively.

Discussion

Our study is pioneering the classification of CGM documents in the EHR, and providing a description of the current status of CGM presentation and storage in clinical practice at a large urban academic medical center. Our algorithm achieved high sensitivity and specificity for CGM AGP classification. Nearly half of the CGM-related documents in the EHR were AGP reports, which are useful for clinical practice and diabetes research; however, the remaining half are other

clinical documents that do not contain useful CGM clinical data. This work is an important step in understanding the management of CGM data in the EHR, which is helpful for the timely delivery of glucose data to providers or the creation of automated reports based on CGM data. The long-term goal of work is to use patient-generated glucose data effectively to improve the quality of glycemic control for patients with diabetes in the current clinical practice.

The AGP report is now established as the standardized, practical report for graphically presenting a summary of glycemic control status in patients with diabetes who use CGM as part of daily diabetes care.[18] The AGP report offers both visual and statistical summaries of glucose metrics, aligning with the 2019 international consensus for evaluating glycemic control, which mandates analysis for all individuals with diabetes utilizing CGM systems.[16] These reports are from the most common products by Abbott Diabetes Care (Abbott Laboratories, Abbott Park, Illinois) and Dexcom (Dexcom, Inc, San Diego, California). They pair the sensor with the sensor's reader app (eg, Abbott FreeStyle Libre Link or Dexcom G6 app), the follower app (eg, LibreLinkUp or Dexcom Follow), and software to review patterns on a computer or mobile device (eg, LibreView or Clarity).[11] None of these applications have standard, replicable integrations with EHR systems that enable passing discrete data captured in the system directly into patient charts. Presently, static images of AGP CGM reports are viewable by clinicians in the EHR, which do not allow for easily finding relevant clinical data or visualizing longitudinal data, although research on automated integration of CGM data in EHR is emerging.[19,20] Currently, in most hospital or clinic settings, rich CGM data being collected on patients are not able to be used to easily provide clinical decision support related to glycemic management.

The main limitation of the study was that we downloaded CGM-related documents from the EPIC system (eg, Media) and then classified these documents. When using this algorithm, it may not be applicable to a system that does not store CGM-related documents in a decentralized location, making it difficult to retrieve these documents in the first step. However, our study is pioneering in the classification of CGM documents stored within the EHR system.

The CGM reports are a powerful tool for the clinical management of diabetes, and they will become more useful as they are integrated into the clinician's workflow and contextualized with the rest of the patient's clinical data. We would eventually want the data to go into the EHR system, similar to a point of care glucose. Future work needs to standardize the storage of CGM-related documents in the EHR. Standardized approaches to transmitting and storing CGM data in the EHR that are easy to search and analyze need to be developed.

Abbreviations

CGM, continuous glucose monitoring; AGP, ambulatory glucose profile; EHR, electronic health record; OCR, optical character recognition.

ORCID iDs

Yaguang Zheng https://orcid.org/0000-0002-8400-1398
Lehan Li https://orcid.org/0009-0000-4590-2081
Zhihao Chen https://orcid.org/0009-0006-1479-7892

References

1. Department of Health and Human Services U.S. centers for disease control and prevention. national diabetes statistics report: prevalence of both diagnosed and undiagnosed diabetes Accessed August 30, 2023. https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html
2. American Diabetes Association. Economic costs of diabetes in the U.S. in 2017. *Diabetes Care*. 2018;41(5):917-928.
3. Boye KS, Thieu VT, Lage MJ, Miller H, Paczkowski R. The association between sustained HbA1c control and long-term complications among individuals with type 2 diabetes: a retrospective study. *Adv Ther*. 2022;39(5):2208-2221.
4. Lafeuille MH, Grittner AM, Gravel J, et al. Quality measure attainment in patients with type 2 diabetes mellitus. *Am J Manag Care*. 2014;20(1 suppl):s5-s15.
5. Bao S, Bailey R, Calhoun P, Beck RW. Effectiveness of continuous glucose monitoring in older adults with type 2 diabetes treated with basal insulin. *Diabetes Technol Ther*. 2022;24(5):299-306.
6. Bristol AA, Litchman M, Berg C, et al. Using continuous glucose monitoring and data sharing to encourage collaboration among older adults with type 1 diabetes and their care partners: qualitative descriptive study. *JMIR Nurs*. 2023;6:e46627.
7. Cox DJ, Oser T, Moncrief M, Conaway M, McCall A. Long-term follow-up of a randomized clinical trial comparing glycemic excursion minimization (GEM) to weight loss (WL) in the management of type 2 diabetes. *BMJ Open Diabetes Res Care*. 2021;9(2):e002403.
8. Kieu A, King J, Govender RD, Östlundh L. The benefits of utilizing continuous glucose monitoring of diabetes mellitus in primary care: a systematic review. *J Diabetes Sci Technol*. 2023;17(3):762-774.
9. Laffel LM, Kanapka LG, Beck RW, et al. Effect of continuous glucose monitoring on glycemic control in adolescents and young adults with type 1 diabetes: a randomized clinical trial. *JAMA*. 2020;323(23):2388-2396.
10. Zipp R. CGM patients seen rising 38% in 2021 fueled by Type 2 diabetes: poll. Published 2021. Accessed May 1, 2023. https://www.medtechdive.com/news/cgm-patients-seen-rising-38-in-2021-fueled-by-type-2-diabetes-poll/596914/
11. Espinoza J, Xu NY, Nguyen KT, Klonoff DC. The need for data standards and implementation policies to integrate CGM

data into the electronic health record. *J Diabetes Sci Technol*. 2023;17(2):495-502.

12. Danne T, Nimri R, Battelino T, et al. International consensus on use of continuous glucose monitoring. *Diabetes Care*. 2017;40(12):1631-1640.

13. Carlson AL, Mullen DM, Bergenstal RM. Clinical use of continuous glucose monitoring in adults with type 2 diabetes. *Diabetes Technol Ther*. 2017;19(S2):S4-S11.

14. Espinoza J, Shah P, Raymond J. Integrating continuous glucose monitor data directly into the electronic health record: proof of concept. *Diabetes Technol Ther*. 2020;22(8):570-576.

15. Lin R, Brown F, Ekinci EI. The ambulatory glucose profile and its interpretation. *Med J Aust*. 2022;217(6):295-298.

16. Battelino T, Danne T, Bergenstal RM, et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes Care*. 2019;42(8):1593-1603.

17. Keras-ocr. Accessed March 1, 2023. https://keras-ocr.readthedocs.io/en/latest/

18. Czupryniak L, Dzida G, Fichna P, et al. Ambulatory Glucose Profile (AGP) report in daily care of patients with diabetes: practical tips and recommendations. *Diabetes Ther*. 2022;13(4):811-821.

19. Kumar RB, Goren ND, Stark DE, Wall DP, Longhurst CA. Automated integration of continuous glucose monitor data in the electronic health record using consumer technology. *J Am Med Inform Assoc*. 2016;23(3):532-537.

20. Aleppo G, Chmiel R, Zurn A, et al. Integration of continuous glucose monitoring data into an electronic health record system: single-center implementation. *J Diabetes Sci Technol*. 2025;19(2):426-430.