

# Availability of Open Dynamic Glycemic Data in the Field of Diabetes Research: A Scoping Review

Journal of Diabetes Science and Technology  
1–13  
© 2025 Diabetes Technology Society  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/19322968251316896  
journals.sagepub.com/home/dst



Libera Lucia Del Giudice, MSc<sup>1</sup> , Agnese Piersanti, PhD<sup>2</sup> ,  
Christian Göbl, MD, PhD, MSc<sup>3</sup>, Laura Burattini, PhD<sup>1</sup> ,  
Andrea Tura, PhD<sup>2</sup>, and Micaela Morettini, PhD<sup>1</sup> 

## Abstract

**Background:** Poor data availability and accessibility characterizing some research areas in biomedicine are still limiting potentialities for increasing knowledge and boosting technological advancement. This phenomenon also characterizes the field of diabetes research, in which glycemic data may serve as a basis for different applications. To overcome this limitation, this review aims to provide a comprehensive analysis of the publicly available data sets related to dynamic glycemic data.

**Methods:** Search was performed in four different sources, namely scientific journals, Google, a comprehensive registry of clinical trials and two electronic databases. Retrieved data sets were analyzed in terms of their main characteristics and on the typology of data provided.

**Results:** Twenty-five data sets were identified including data from challenge tests (5 of 25) or data from Continuous Glucose Monitoring (CGM, 20 of 25). As for the data sets including challenge tests, all of them were freely downloadable; most of them (80%) related only to oral glucose tolerance test (OGTT) with standard duration (2 h), but varying for timing and number of collected blood samples, and variables collected in addition to glucose levels (with insulin levels being the most common); the remaining 20% of them also included intravenous glucose tolerance test (IVGTT) data. As for the data sets related to CGM, 7 of 20 were freely downloadable, whereas the remaining 13 were downloadable upon completion of a request form.

**Conclusions:** This review provided an overview of the readily usable data sets, thus representing a step forward in fostering data access in diabetes field.

## Keywords

artificial intelligence, continuous glucose monitoring, data set, diabetes, glycemia, repository

## Introduction

The so-called “data tsunami” phenomenon—being defined as the possibility to generate data from multiple sources—is increasingly impacting several fields and in particular the health care field.<sup>1</sup> However, this potential data availability does not always translate into an authentic opportunity for data access; thus, fostering the open data has become a priority in policy initiatives, especially to potentiate the impact of the current artificial intelligence (AI) solutions.<sup>2</sup>

Diabetes care is one of the health care areas for which the data tsunami may have a considerable impact. As an example, it is worth mentioning the great amount of data generated from wearable devices, such as continuous glucose monitoring (CGM) devices, insulin pumps, or other devices, which can be incorporated into electronic health records and exploited in digital health technologies like AI or telehealth solutions.<sup>3</sup> However, at the same time, this area suffers

remarkably from the lack of available data,<sup>4</sup> thus limiting the potentialities for technological advancements in personalized patient care.<sup>5</sup> This may be due to multiple reasons connected to the nature of the relevant data and the issues in producing and sharing them. First, data acquisition involves sampling of blood or interstitial fluid to quantify metabolites and/or hormone concentrations, thus implying invasive or minimally

<sup>1</sup>Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

<sup>2</sup>CNR Institute of Neuroscience, Padua, Italy

<sup>3</sup>Division of Obstetrics and Feto-Maternal Medicine, Department of Obstetrics and Gynaecology, Medical University of Vienna, Vienna, Austria

### Corresponding Author:

Micaela Morettini, PhD, Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 12, Ancona 60131, Italy.

Email: m.morettini@univpm.it

invasive procedures performed only in clinical settings or under the supervision of medical personnel. Second, data sets including several variables are usually collected through procedures that are demanding in terms of time and/or money, and for this reason, they are performed only during clinical trials rather than in routine clinical practice. This often translates into data collection on a limited number of subjects. Finally, data sets from routine clinical practice can be large in terms of number of subjects, but they typically include a limited number of variables and they are difficult to be accessed for administrative issues. In order to support research on diabetes and help data retrieval, this review aimed at conducting a comprehensive search to locate and categorize data sets containing glycemetic data that are accessible to the public. In particular, this review focused on dynamic glycemetic data, which are of interest for the development of many technological applications.<sup>6</sup>

## Methods

### Eligibility Criteria

This review targeted scientific articles, websites, and clinical trials that provide free access to data sets including dynamic glycemetic data (both human and animal). Both immediately downloadable data sets and data sets downloadable after completion of a request form were included.

### Exclusion Criteria

This analysis excluded scientific articles, websites, and clinical trials: (1) that do not contain any type of data set; (2) that contain an unavailable data set or a data set available upon payment of subscription fees; (3) in which the data of interest are mainly expressed as a fasting glucose or glycated hemoglobin A<sub>1c</sub> (HbA<sub>1c</sub>) as a single value, being very often collected within a basic routine examination and also as secondary variables (ie, belonging to a data set whose primary aim is not collecting glycemetic data useful for diabetes-related research). This latter criterion was necessary to exclude data sets with some diabetes-related information but with limited usefulness for novel diabetes-related research, being them likely already exploited in the original studies for which they were collected.

### Literature Search Strategy

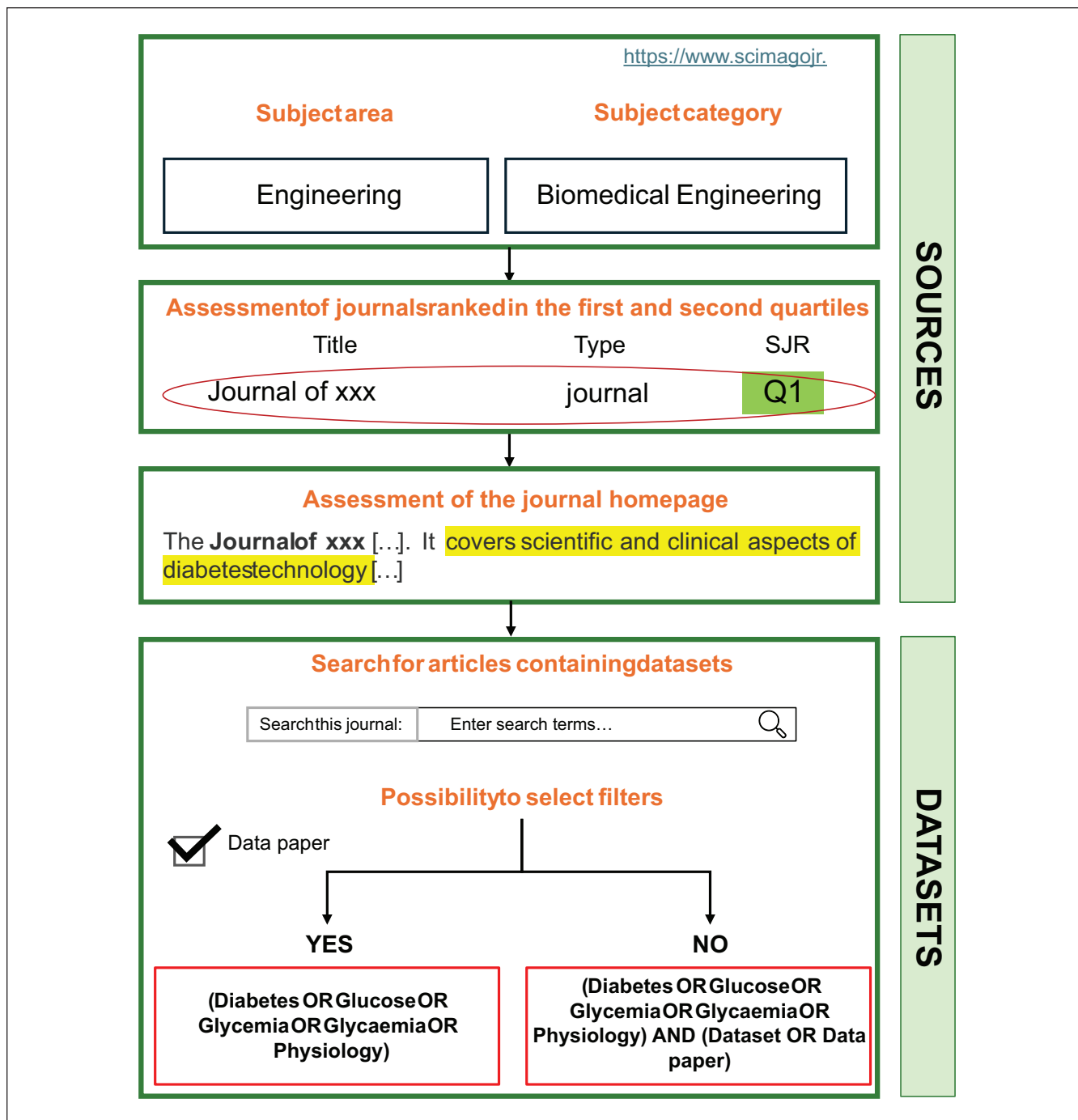
This scoping review was conducted according to the Arksey and O'Malley methodological framework<sup>7</sup> and the guidelines provided by Daut et al<sup>8</sup> and Peters et al.<sup>9</sup> The PRISMA extension for Scoping Reviews (PRISMA-ScR) completed checklist<sup>10</sup> was provided as Supplementary Material. To discover all the available data sets, a systematic search of scientific articles, websites, and clinical trials was carried out utilizing four different sources, namely: (1) high-ranking scientific journals in the topic domain of interest, (2) Google,

(3) a comprehensive clinical trial registry, and (4) two electronic databases. To perform the search, two groups of terms were considered and combined; in details, the first group included the terms referring to the resource of interest (eg, "Repository," "Biobank," "Platform," and "Dataset") and the second group included the terms concerning the topics domain of interest (eg, "Diabetes," "Glucose," "Glycemia," "Glycaemia," "Physiology"). Details on the Boolean logic operators used to combine the various keywords and on the strategy applied for each of the four sources are provided in the following subsections. The search was initially conducted in the period from February 2023 to May 2023, but the last update to search was performed in December 2024. The English language was set as a filter.

**Scientific journal search.** High-ranking scientific journals were selected according to the publicly available portal Scimago Journal & Country Rank (SJR).<sup>11</sup> Scientific journals appropriate for this research were selected by considering "Medicine" as subject area and "Endocrinology, Diabetes and Metabolism" as subject category or considering "Engineering" as subject area and "Biomedical engineering" as subject category. Only journals ranked in the first two quartiles (Q1, Q2) were considered. For each selected journal, its homepage was accessed and a second search for the relevant journal articles was performed. For journals providing the opportunity to select the type of results (ie, insert the filter "Dataset" or "Data paper") directly on their homepage, a second search was performed using the terms of the second group connected with the comma or the "OR" (the choice depending on the portal instructions). If this option was not available, a search was performed by linking through the "AND" operator all the terms of the second group (connected with "OR") and the search string ("Dataset" OR "Data Paper"). Steps for the scientific journal search strategy are summarized in Figure 1.

**Google search.** Five different searches of websites were performed according to the following strategy: for each search, all the terms belonging to the first group were considered and only one of the five terms of the second group was chosen; the terms within the first group were linked by the Boolean operator "OR" and then combined with the Boolean operator "AND" with the term chosen in the second group (eg, "Repository OR Biobank OR Platform OR Dataset" AND "Diabetes"). For each search, the first 50 websites ranked in order of relevance were evaluated in relation to the typology of their content (ie, data set, paper with associated data set, repository/portal, etc) similarly to what indicated as the last step of the scientific journal search. Steps for the Google search strategy are summarized in Figure 2.

**Clinical trials search.** The considered clinical trial registry was *ClinicalTrials.gov* (<https://www.clinicaltrials.gov/>). The domain "Condition of Disease" was set to "Diabetes" and in the field "Other Terms" the words "glucose, glycaemia, insulin" were



**Figure 1.** Overview of the steps for the search and screening strategy in scientific journals.

inserted to indicate glycemc data and those data mostly related to it, thus usually collected in diabetes-related studies. The filter “Study with results” was applied. Steps for the search strategy in *ClinicalTrials.gov* are summarized in Figure 3.

**Electronic databases search.** The SCOPUS and PubMed were selected as electronic databases, and the advanced search strategy was implemented. The following search query was applied: (“Diabetes” OR “Glucose” OR “Glycemia”

OR “Physiology”) AND (“Repository” OR “Biobank” OR “Platform” OR “Dataset”). The filters “data paper” or “associated dataset” were applied depending on the database’s instructions.

**Screening Strategy**

For each typology of search, a two-level screening strategy was performed as detailed in the following.

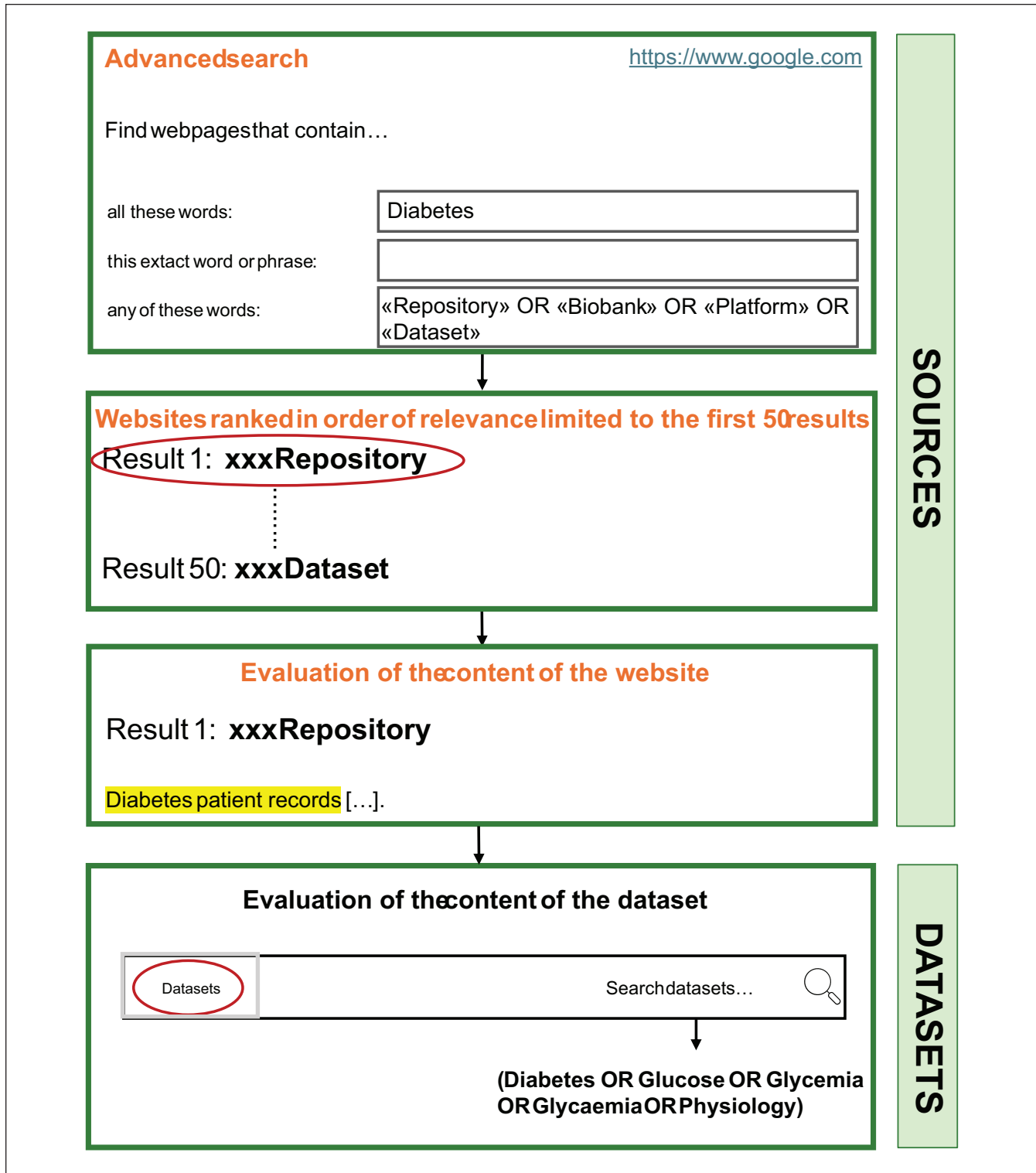
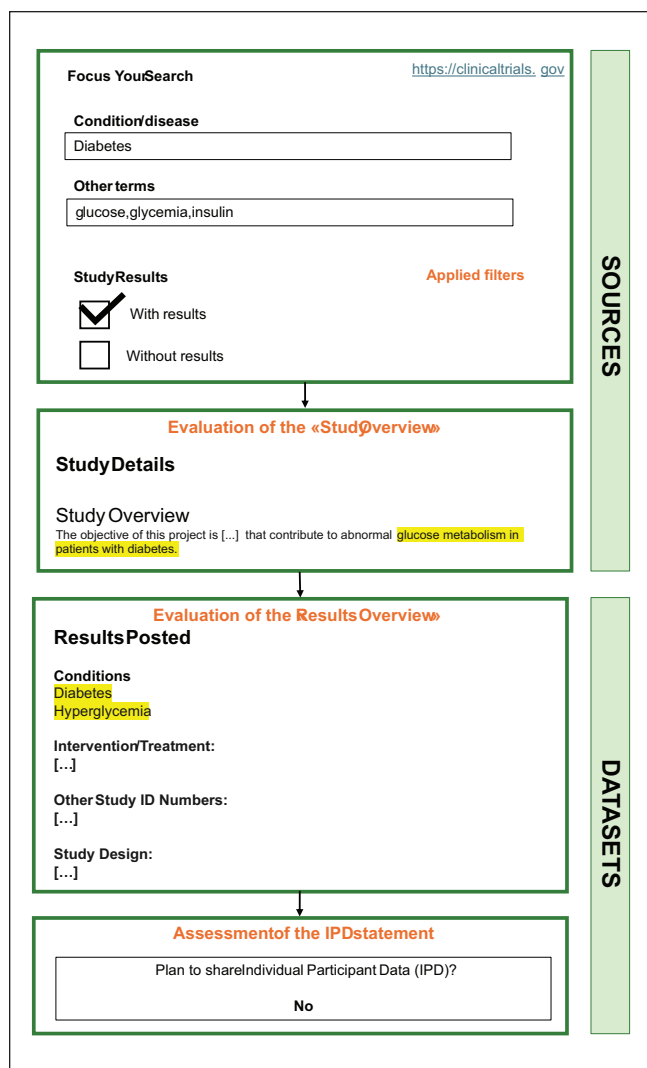


Figure 2. Overview of the steps for the search and screening strategy in Google.



**Figure 3.** Overview of the steps for the search and screening strategy in the clinical trial registry ClinicalTrials.gov.

**Screening strategy for journal search.** For each selected journal, the journal's scope was first screened; if the journal was retained based on its scope, a second screening was performed at the article level considering also supplementary materials when applicable, according to eligibility/exclusion criteria detailed above (Figure 1).

**Screening strategy for Google search.** The 50 websites resulting for each of the five searches were first screened without fully accessing them. If the website was deemed appropriate, a more advanced screening was performed with detailed access of the website (Figure 2).

**Screening strategy for clinical trials.** Screening for the clinical trials was performed first by details reported in "Study overview" and then by those reported in "Results overview" (Figure 3).

**Screening strategy for electronic databases.** After removing duplicates between the two sources, all the record titles were assessed as a first screening step. Then, a second screening was performed at the full-text and data set level.

## Data Analysis

Each data set was analyzed in terms of the publication year, the study design, the population involved (humans or animals), the number of subjects participating in the study, information on gender and age, the type of diabetes, the type of dynamic glyceamic data (ie, oral or intravenous glucose tolerance test [IVGTT], CGM), the test duration, the number of acquired samples and variables, or, if applicable, the devices employed for the measurement and the condition of the study (laboratory/hospital or free-living conditions).

## Results

A total of 25 data sets were included in the review, following screening and selection as detailed in the flowchart in Figure 4. The portals in which the data sets are located are listed in Table 1. Retrieved data sets included two main typologies of dynamic glyceamic data, namely data from challenge tests (5 data sets)<sup>12-16</sup> and data from CGM (20 data sets).<sup>17-36</sup> As for the first typology, data were related to oral (oral glucose tolerance test [OGTT]) and, in a small number of cases (1 of 5) also to IVGTT, as detailed in Table 2. The duration of the OGTT in the considered data sets is equal to 2 hours, whereas the number of blood samples acquired showed variability among the data sets. Also, the number of variables acquired during the test can vary depending on the objectives of the study, as it is shown in Table 3; a common characteristic in the majority of the challenge test data sets (3 of 5) is the assessment of both glyceamic and insulinemia. All the challenge test data sets are freely available and downloadable.

The second typology of data pertains to CGM, which represents a technique to monitor glucose every 1 to 5 minutes for 7 to 10 days or even more with a single glucose sensor.<sup>37</sup> As regards CGM data, it is necessary to distinguish between data sets that are freely available and directly downloadable from any user (detailed in Table 4) and data sets that are available under completion of a request form (detailed in Table 5). Such data sets are characterized by different duration, spanning from acquisitions lasting less than two weeks to acquisitions lasting more than nine weeks (Figure 5). Moreover, the number of subjects in the CGM data sets can vary, with the majority (11 of 20) characterized by more than 100 subjects (Figure 6).

## Discussion

This review systematically analyzed the freely available data sets containing dynamic glyceamic data that can be exploited in diabetes research. The 25 data sets included

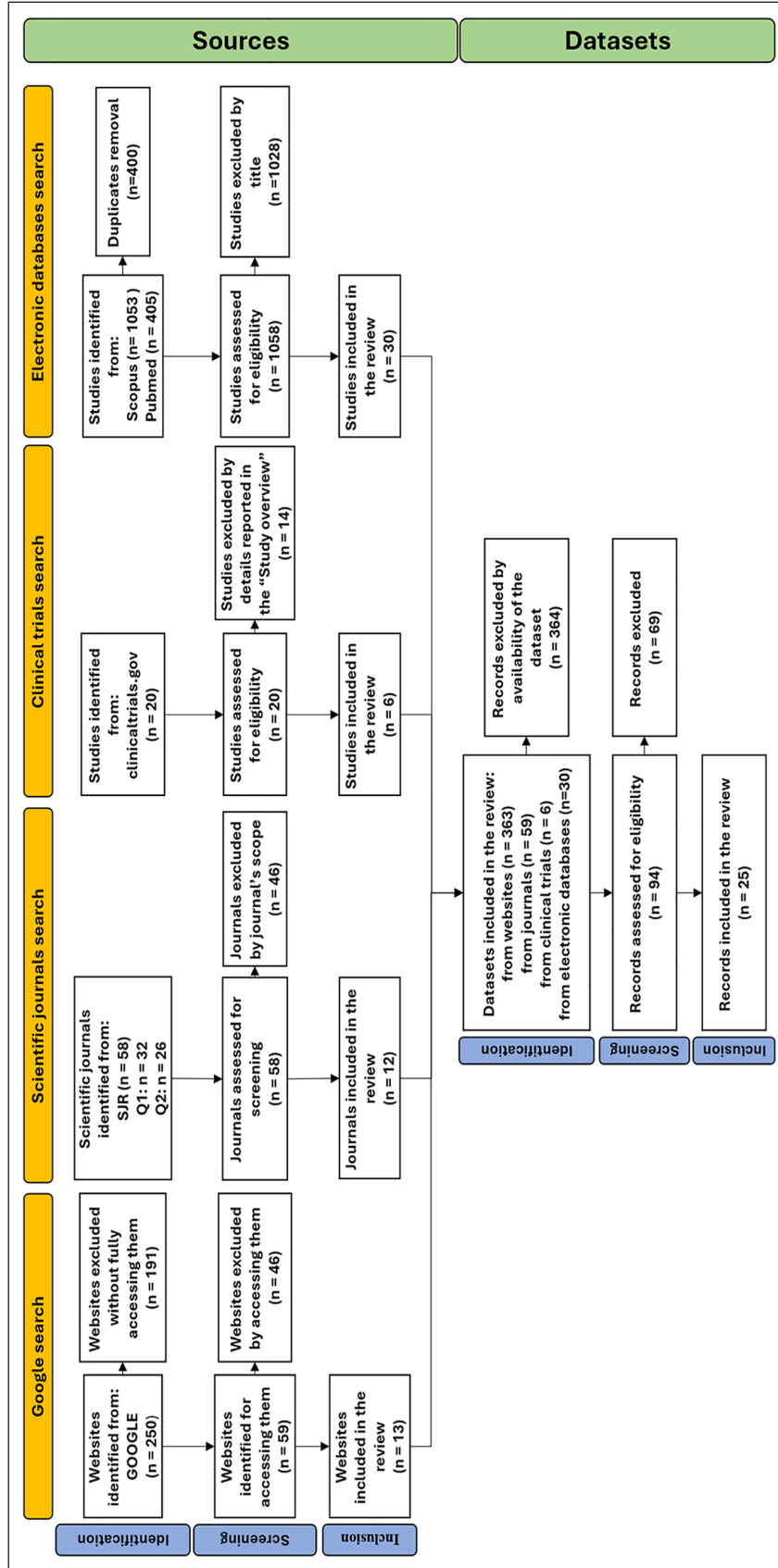


Figure 4. PRISMA flowchart.

were retrieved from four different searches (ie, scientific journals, Google, clinical trials, and electronic databases), which made us confident that all potential sources of relevant information were covered. As regards the scientific journals search, a one-by-one screening was applied for scientific journals in the field, limiting the search to those in the top quartiles (Q1 and Q2, as per SJR). Although this choice may expose to the risk of not considering data sets

potentially published in Q3-Q4 quartile journals, it was considered a reasonable choice since the most relevant quartiles were included. Moreover, a standard search in the databases of scientific literature (like SCOPUS and PubMed) was also included to complement the plethora of data papers or scientific papers with associated data available. In SCOPUS, the application of the filter “data paper” was necessary since the standard search was not viable due to the impossibility of applying a criterion to filter articles with associated available data, which is instead possible in PubMed. Of note, the strategy we applied for journal searching allows also to not filter a priori relevant results published as abstract only, if present in high-ranking scientific journals. The inclusion of Google search, as well as search in clinical trials, was motivated by the willingness to identify suitable websites and online sources where relevant data sets could potentially be located, even if they may not have an associated paper in a scientific journal. The exhaustive overview of the main websites and online sources provided by our search process is demonstrated by results reported in Table 1, which include, to the best of our knowledge, all the main portals devoted to data sharing. However, the implementation of a preliminary screening for the website search was necessary since Google is a vast search engine, housing an immense

**Table 1.** List of the Portals in Which the Data Sets are Located.

Name of the portal	Link to the portal
zenodo	<a href="https://zenodo.org">https://zenodo.org</a>
github	<a href="https://github.com">https://github.com</a>
kaggle	<a href="https://kaggle.com">https://kaggle.com</a>
figshare	<a href="https://figshare.com">https://figshare.com</a>
data.world	<a href="https://data.world">https://data.world</a>
ResearchGate	<a href="https://researchgate.net">https://researchgate.net</a>
DRYAD	<a href="https://datadryad.org">https://datadryad.org</a>
PhysioNet	<a href="https://physionet.org">https://physionet.org</a>
Mendeley	<a href="https://mendeley.com">https://mendeley.com</a>
IEEEDataPort	<a href="https://iee-dataport.org">https://iee-dataport.org</a>
JAEB	<a href="https://www.jaeb.org">https://www.jaeb.org</a>
DATA.GOV	<a href="https://data.gov">https://data.gov</a>

**Table 2.** Summary of the Data Sets Related to Challenge Tests.

Ref.	Data sets' name/ authors (link to access)	Year	Study design	Type of population	No. of subjects	Sex (M/F)	Age (years)	Type of diabetes	Challenge test (duration)	No. of samples
12	Pima Indians Diabetes Database ( <a href="https://kaggle.com/datasets/uciml/pima-indians-diabetes-database">https://kaggle.com/datasets/uciml/pima-indians-diabetes-database</a> )	–	O	Human	768	F	21	–	OGTT (2 h)	1
13	Edinburgh et al. ( <a href="https://researchdata.bath.ac.uk/352">https://researchdata.bath.ac.uk/352</a> )	2017	I	Human	10	M	–	Healthy	OGTT (2 h)	7
14	Manell et al. ( <a href="https://datadryad.org/stash/dataset/doi:10.5061/dryad.n3f4j">https://datadryad.org/stash/dataset/doi:10.5061/dryad.n3f4j</a> )	2017	–	Animal (Pigs)	18	9M/9F	–	Healthy	OGTT (2 h) IVGTT (3 h)	11
15	Tauer et al. ( <a href="https://data.mendeley.com/datasets/np7kpnk9t4/1">https://data.mendeley.com/datasets/np7kpnk9t4/1</a> )	2021	I	Animal (Mice)	139	70M/69F	–	–	OGTT (2 h)	6
16	Flores-Arguedas et al. ( <a href="https://github.com/hugofloresar/OGTT/blob/main/Datos_OGTT.xlsx">https://github.com/hugofloresar/OGTT/blob/main/Datos_OGTT.xlsx</a> )	2021	O	Human	52	F	–	–	OGTT (2 h)	5

Study design: I (interventional), O (observational); when possible, age is expressed as mean  $\pm$  standard deviation; M/F: male/female; type of diabetes (type 1, type 2, healthy, non-diabetic); “–” indicates that the information is not provided in the related article/data set description or does not match between the description and what is effectively found in the data set.

**Table 3.** Variables Measured During the Challenge Test and Reported in the Data Sets.

Ref.	Variables						
	Glucose	Insulin	Glucagon	C-peptide	GLP-I	Triglyceride	Lactate
12	x	x					
13	x	x				x	x
14	x	x	x		x		
15	x						
16	x						

GLP-I: glucagon like peptide-1.

**Table 4.** Summary of the Freely Available and Downloadable Data Sets Related to Continuous Glucose Monitoring.

Ref.	Data sets' name/ authors (website)	Year	Study design	Type of population	No. of subjects	Sex (M/F)	Age (years)	Type of diabetes	Condition	Device	Duration
17	Hidalgo et al. ( <a href="https://data.mendeley.com/datasets/3hbcscwz44/1">https://data.mendeley.com/datasets/3hbcscwz44/1</a> )	2024	O	Human	25	12M/13F	40.50 ± 11.43 37.63 ± 12.80	Type 1	Free-living	Abbot; FreeStyle Libre 2	14 days
18	Zhao et al. ( <a href="https://figshare.com/collections/Diabetes_Datasets_ShanghaiT1DM_and_ShanghaiT2DM/6310860/2">https://figshare.com/collections/Diabetes_Datasets_ShanghaiT1DM_and_ShanghaiT2DM/6310860/2</a> )	2023	I	Human	12 100	5M/7F 56M/44F	57.8 ± 11.1 60.2 ± 13.7	Type 1 Type 2	-	Abbot; FreeStyle Libre	14 days
19	Colás et al. ( <a href="https://figshare.com/articles/dataset/Detrended_Fluctuation_Analysis_in_the_prediction_of_type_2_diabetes_mellitus_in_patients_at_risk_Model_optimization_and_comparison_with_other_metrics/11398914">https://figshare.com/articles/dataset/Detrended_Fluctuation_Analysis_in_the_prediction_of_type_2_diabetes_mellitus_in_patients_at_risk_Model_optimization_and_comparison_with_other_metrics/11398914</a> )	2019	I	Human	208	-	18+	Type 2	Free-living	Medtronic; iPro	24 h
20	Åm et al. ( <a href="https://datadryad.org/stash/dataset/doi:10.5061/dryad.5m1m755">https://datadryad.org/stash/dataset/doi:10.5061/dryad.5m1m755</a> )	2018	-	Animal (Pigs)	12	-	-	Non-diabetic	Lab	Abbot; FreeStyle Libre	8 h
21	Tamborlane et al. ( <a href="https://github.com/IrinaStatsLab/Awesome-CGM/wiki/Tamborlane-(2008)">https://github.com/IrinaStatsLab/Awesome-CGM/wiki/Tamborlane-(2008)</a> )	2008	I	Human	322 129	-	-	Type 1	Free-living	Abbot; FreeStyle Navigator, Dexcom; SEVEN, Medtronic; Paradigm	6 months
22	Dubosson et al. ( <a href="https://zenodo.org/records/5651217">https://zenodo.org/records/5651217</a> )	2018	O	Human	9	-	18+	Type 1	Free-living	Medtronic; iPro2	4 days
23	Hall et al. ( <a href="https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2005143#pbio.2005143.s010">https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2005143#pbio.2005143.s010</a> )	2018	I	Human	57	25M/32F	25-76	Non-diabetic	Free-living	Dexcom; G4	2-4 weeks

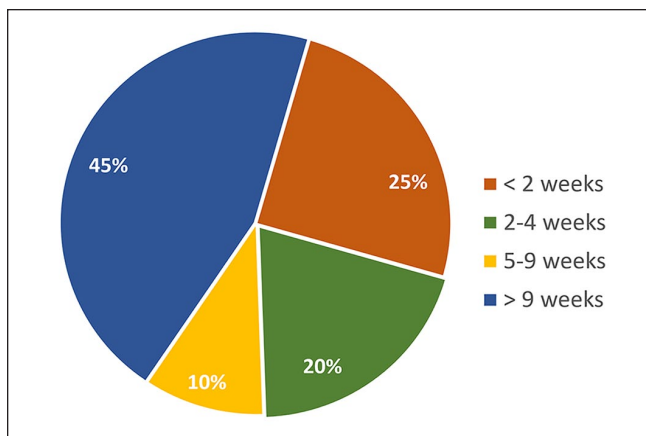
Study design: I (interventional), O (observational); when possible, age is expressed as mean ± standard deviation; M/F: male/female; type of diabetes (type 1, type 2, healthy, non-diabetic); condition refers to that in which the study was performed (free-living condition or in a lab/hospital); device refers to the CGM device used; “-” indicates that the information is not provided in the related article/data set description or does not match between the description and what is effectively found in the data set.



**Table 5.** Summary of the Data Sets Related to Continuous Glucose Monitoring, Freely Available Upon Completion of a Request Form.

Ref.	Data Sets' name/authors (website)	Year	Study design	Type of population	No. of subjects	Sex (M/F)	Age (years)	Type of diabetes	Condition	Device	Duration
24	Aleppo et al. ( <a href="https://github.com/irinagain/Awesome-CGM/wiki/Aleppo-(2017)">https://github.com/irinagain/Awesome-CGM/wiki/Aleppo-(2017)</a> )	2017	I	Human	225	-	25-40	Type I	Free-living	Dexcom; G4	6 months
25	Anderson et al. ( <a href="https://public.jaeb.org/drfapp2/study/465">https://public.jaeb.org/drfapp2/study/465</a> )	2015	I	Human	30	-	18+	Type I	Free-living	Dexcom; G4 Platinum	6-9 weeks
26	Weinstock et al. ( <a href="https://github.com/irinagain/Awesome-CGM/wiki/Weinstock-(2016)">https://github.com/irinagain/Awesome-CGM/wiki/Weinstock-(2016)</a> )	2016	CC	Human	200	-	60+	Type I	Free-living	Dexcom; SEVEN PLUS	14 days
27	Buckingham et al. ( <a href="https://public.jaeb.org/dir-ecnet/study/166">https://public.jaeb.org/dir-ecnet/study/166</a> )	2006	I	Human	30	-	3-17	Type I	Free-living	Abbot; FreeStyle Navigator	13 weeks
28	Chase et al. ( <a href="https://public.jaeb.org/dir-ecnet/study/159">https://public.jaeb.org/dir-ecnet/study/159</a> )	2004	I	Human	200	-	7-18	Type I	Free-living	Cygnus; GlucoWatch G2 Biographer (GWB)	6 months
29	Tsalikian et al. ( <a href="https://public.jaeb.org/dir-ecnet/study/160">https://public.jaeb.org/dir-ecnet/study/160</a> )	2005	I	Human	50	-	10-18	Type I	Hospital	OneTouch Ultra Meter	48 h
30	Ohio T1DM ( <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7881904/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7881904/</a> )	2021	O	Human	12	-	-	Type I	Free-living	Medtronic; Enlite	8 weeks
31	The Pediatric Artificial Pancreas (PEDAP) Trial of Control-IQ Technology in Young Children in Type I Diabetes ( <a href="https://public.jaeb.org/dataset/599">https://public.jaeb.org/dataset/599</a> )	2024	I	Human	150	-	2-6	Type I	Free-living	Dexcom; G5 Dexcom; G6	26 weeks-7 months
32	Hybrid Closed Loop Therapy and Verapamil for Beta Cell Preservation in New Onset Type I Diabetes (CLVer) ( <a href="https://public.jaeb.org/dataset/591">https://public.jaeb.org/dataset/591</a> )	2022	I	Human	98 33	-	7-18	Type I	Free-living	Dexcom; G6	12 months
33	A Study to Assess Continuous Glucose Sensor Profiles in Healthy Non-Diabetic Participants Aged <7 Years ( <a href="https://public.jaeb.org/dataset/593">https://public.jaeb.org/dataset/593</a> )	2021	O	Human	50	-	1-7	Non-diabetic	Free-living	Dexcom; G6	10 days
34	FLAIR- Fuzzy Logic Automated Insulin Regulation: A Crossover Study Comparing Two Automated Insulin Delivery System Algorithms (PID vs. PID + Fuzzy Logic) in Individuals with Type I Diabetes ( <a href="https://public.jaeb.org/dataset/566">https://public.jaeb.org/dataset/566</a> )	2020	I	Human	113	43M/70F	19 ± 4	Type I	Free-living	Medtronic; MiniMed Guardian	28-36 weeks
35	The International Diabetes Closed Loop (iDCL) trial: Clinical Acceptance of the Artificial Pancreas—A Pivotal Study of t:slim X2 with Control-IQ Technology (DCLP3) ( <a href="https://public.jaeb.org/dataset/573">https://public.jaeb.org/dataset/573</a> )	2022	I	Human	168	84M/84F	33 ± 16	Type I	Free-living	Dexcom; G6	6 months
36	CGM Intervention in Teens and Young Adults with T1D (CITY): A Randomized Clinical Trial to Assess the Efficacy and Safety of Continuous Glucose Monitoring in Young Adults 14-<25 with Type I Diabetes ( <a href="https://public.jaeb.org/dataset/565">https://public.jaeb.org/dataset/565</a> )	2019	I	Human	150	-	14-25	Type I	Free-living	Dexcom; G5	6 months

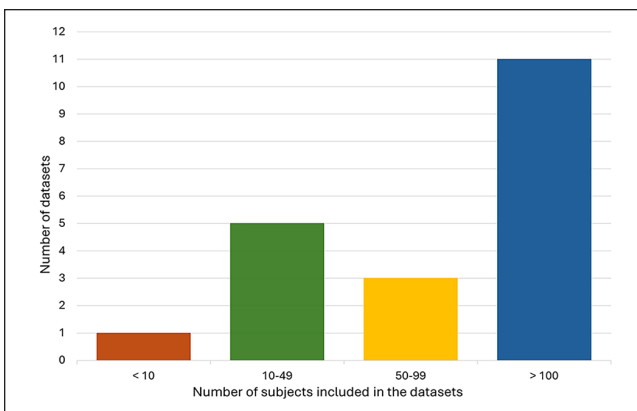
Study design: I (interventional); O (observational); CC (case-control); when possible, age is expressed as mean ± standard deviation; M/F: male/female; type of diabetes (type 1, type 2, healthy, non-diabetic); condition refers to that in which the study was performed (free-living condition or in a lab/hospital); device refers to the CGM device used; “-” indicates that the information is not provided in the related article/data set description or does not match between the description and what is effectively found in the data set.



**Figure 5.** Duration of the continuous glucose monitoring (CGM) acquisitions across the reviewed data sets.

volume of data and, without some form of filtering, the provided results can become overwhelming and require substantial time for examination. As regards the clinical trials search, although different clinical trial registries could be considered, the search was limited to *ClinicalTrials.gov*; however, we did not consider this as a substantial limitation, as this registry is one of the largest and most used.

Overall, the outcomes derived from this analysis indicate that the available dynamic glycemic data can be classified into two main categories: challenge test and CGM data. As for the first category, all data from challenge tests pertain to OGTT. This result is not unexpected, being the 75-g OGTT one of the diagnostic screening tests for diabetes and prediabetes according to the American Diabetes Association (ADA) guidelines.<sup>38</sup> Concerning the variables provided within the OGTT data sets, the majority of the data sets reported the measurement of additional variables with respect to glycemia; in particular, insulin levels feature prominently in most of the data sets, whereas other information such as C-peptide, triglycerides, and glucagon-like peptide-1 (as a marker of incretin action) are rarely measured. It is also worth noting that all the OGTTs have standard duration (2 h). Conversely, with regard to the number of OGTT blood samples, it is noteworthy that the greater majority of the retrieved data sets provide a number of samples higher than that required for diabetes diagnosis (ie, two blood samples, at fasting and at 2 h during OGTT). Indeed, the collection of four or five samples at 30-minute intervals is confirmed as a common approach,<sup>39</sup> with some study protocols collecting even more samples. As regards IVGTT, only one of the data sets reported this type of data. Indeed, the IVGTT is a test typically used in the investigation of diabetes pathophysiology<sup>40</sup> and hence performed for research purposes rather than for direct clinical applications, thus usually resulting in data sets with a limited number of subjects. Of note, this review also targeted data from preclinical in vivo models, frequently studied with OGTT and IVGTT in diabetes pathophysiology research.<sup>41</sup>



**Figure 6.** Number of subjects included in the continuous glucose monitoring (CGM) data sets.

While accounting for a small percentage of data sets included in our review, it is worth noting that these data sets are characterized by a higher number of variables than those provided by studies on humans.

Applications exploiting this category of data may vary not only in relation to the number of subjects included in the data set but also to the protocol used for data acquisition. One of the most consolidated applications relies on the use of challenge test data jointly with mathematical (compartmental) models to extract parameters of clear physiological meaning (eg, insulin sensitivity, alpha and beta-cell sensitivity).<sup>42-47</sup> It is worth noting that for this kind of application, the crucial factor is not represented by the number of subjects in the data set (being analysis performed on an individual basis) but rather by the characteristics of the protocol, which may lack suitability if the number of blood samples and/or variety of data do not match with the requirements in terms of model parameters to be estimated. An illustrative instance of this is the standard OGTT exclusively measuring glucose levels, which provides data not adequate for model-based approaches. Furthermore, other applications may rely on AI-based models, which have more urgent requirements in terms of the number of subjects included in the data set rather than protocol characteristics. Indeed, studies on open and proprietary data sets showed that, when a sufficiently large amount of data are exploitable (ie, in terms of number of subjects), even the basic challenge test data acquired in clinical practice, like the standard OGTT, may become useful to develop AI solutions with real applicability.<sup>48</sup> Also, when the exploitation of mathematical models is enabled by protocol characteristics, model-based features with clear physiological meaning can be extracted to power information gathered from the data and to feed AI models.<sup>49</sup>

The CGM data sets cover a consistent percentage of the identified data sets (20 of 25). The use of CGM devices has become increasingly prevalent in recent years, primarily due to significant advancements in technology. By providing real-time and continuous monitoring of glucose levels (every

1-5 min), these devices are able to provide a large amount of data. Besides, an even more important advantage is their capability to monitor and describe the glucose fluctuations that take place in free-living conditions and in response to perturbations like meals and physical exercise. Information coming from CGM patterns is still widely unexplored, but efforts are in due course to provide standardization in their analysis in relation to the computable CGM metrics.<sup>50</sup> Thanks to the increasing availability of such kind of data sets and to their dimension (most of the reviewed data sets are characterized by more than 100 subjects), information from CGM devices could become crucial for the development of clinical decision support systems in diabetes field based on AI and machine learning approaches. In relation to this aspect and in consideration of the peculiarities of diabetes research field, best practices and pitfalls were described.<sup>4</sup> Finally, it should be acknowledged that the CGM data sets analyzed in this review are mainly related to T1D patients and less frequently to T2D (none in gestational diabetes mellitus [GDM]). On the contrary, this result was somehow expected, being CGM use still typically limited to patients with T1D, even though use in other populations is increasing.<sup>51</sup>

One may argue on the usefulness of such a review study, given the widespread use of generative AI tools also in the field of literature search. However, when we attempt to use generative AI as a surrogate of the literature review here performed (question: “*please search for open data sets containing glucose measurement data which are downloadable freely and put the results in a table*”), only 4 of the 25 data sets were effectively retrieved, which are the most widely known, also easily identifiable as a simple Google search. This implies that, despite being very powerful tools in asking for more information regarding a specific data set, they are not able to replicate the manual effort done in the present study. A second element of criticism can be found in the choice of excluding data sets in which the data of interest are mainly expressed as a fasting glucose or glycated hemoglobin as a single value. The high number of data sets that may have this characteristic, however, would have not represented a real added value for this review; indeed, their potential for studies other than the ones for which they were acquired is usually limited with respect to dynamic data. Eventually, an additional element of criticism can be related to the obsolescence of results here reported, due to the possible availability of open data sets in the future. However, the methodology used in this study has to be considered as a search pipeline, which could be periodically updated to retrieve new data sets. This further emphasizes the usefulness of the present review.

## Conclusions

This review identified a total of 25 data sets that can be freely downloaded. This number represents a small percentage with

respect to the data sets initially anticipated from the four searches performed, proving that poor data accessibility still remains a limitation to overcome in this field. However, the possibility provided by this analysis to have an overview of the readily usable data sets and to easily locate them represents a step forward in fostering data access.

## Abbreviations

ADA, American Diabetes Association; AI, artificial intelligence; CGM, continuous glucose monitoring; GDM, gestational diabetes mellitus; HbA<sub>1c</sub>, glycated hemoglobin; IVGTT, intravenous glucose tolerance test; OGTT, oral glucose tolerance test; PRISMA, Preferred Reporting Items for Systematic Review and Meta-Analysis; T1D, type 1 diabetes; T2D, type 2 diabetes.

## Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding


The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iDs

Libera Lucia Del Giudice  <https://orcid.org/0009-0008-0723-1289>

Agnese Piersanti  <https://orcid.org/0000-0002-1921-838X>

Laura Burattini  <https://orcid.org/0000-0002-9474-7046>

Micaela Morettini  <https://orcid.org/0000-0002-8327-8379>

## Supplemental Material

Supplemental material for this article is available online.

## References

1. Muntz DS. Digital health: how to govern during a never-ending data tsunami. *NPJ Digit Med.* 2021;4:1-3. doi:10.1038/s41746-021-00491-8.
2. Open data and AI: a symbiotic relationship for progress. Accessed January 27, 2025. <https://data.europa.eu/en/publications/datastories/open-data-and-ai-symbiotic-relationship-progress>
3. Klonoff AN, Andy Lee WA, Xu NY, Nguyen KT, DuBord A, Kerr D. Six digital health technologies that will transform diabetes. *J Diabetes Sci Technol.* 2023;17:239-249. doi:10.1177/19322968211043498.
4. Jacobs PG, Herrero P, Facchinetti A, et al. Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls, and opportunities. *IEEE Rev Biomed Eng.* 2024;17:19-41. doi:10.1109/RBME.2023.3331297.
5. Mackenzie SC, Sainsbury CAR, Wake DJ. Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges. *Diabetologia.* 2024;67(2):223-235. doi:10.1007/s00125-023-06038-8.
6. Woldaregay AZ, Årsand E, Walderhaug S, et al. Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes. *Artif Intell Med.* 2019;98:109-134. doi:10.1016/j.artmed.2019.07.007.

7. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8:19-32. doi:10.1080/1364557032000119616.
8. Daudt HML, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Med Res Methodol*. 2013;13:48. doi:10.1186/1471-2288-13-48.
9. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBIM Evid Synth*. 2020;18:2119-2126. doi:10.11124/JBIES-20-00167.
10. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169:467-473. doi:10.7326/M18-0850.
11. Scimago Journal & Country Rank. Accessed January 27, 2025. <https://www.scimagojr.com/>
12. Pima Indians Diabetes Database. Accessed November 10, 2024. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
13. Edinburgh RM, Hengist A, Smith HA, et al. Prior exercise alters the difference between arterialised and venous glycaemia: implications for blood sampling procedures. *Br J Nutr*. 2017;117(10):1414-1421. doi:10.1017/S0007114517001362.
14. Manell E, Hedenqvist P, Svensson A, Jensen-Waern M. Establishment of a refined oral glucose tolerance test in pigs, and assessment of insulin, glucagon and glucagon-like peptide-1 responses. *PLoS ONE*. 2016;11(2):e0148896. doi:10.1371/journal.pone.0148896.
15. Tauer JT, Boraschi-Diaz I, Komarova SV. Data on body mass, glucose tolerance and bone phenotype of mice with osteogenesis imperfecta on long-term low-fat and high-fat diets. *Data Brief*. 2022;41:107961. doi:10.1016/j.dib.2022.107961.
16. Flores-Arguedas H, Capistrán MA, Flores-Arguedas H, Capistrán MA. Bayesian analysis of glucose dynamics during the oral glucose tolerance test (OGTT). *Math Biosci Eng*. 2021;18:4628-4647. doi:10.3934/mbe.2021235.
17. Hidalgo JI, Alvarado J, Botella M, Aramendi A, Velasco JM, Garnica O. HUPA-UCM diabetes dataset. *Data Brief*. 2024;55:110559. doi:10.1016/j.dib.2024.110559.
18. Zhao Q, Zhu J, Shen X, et al. Chinese diabetes datasets for data-driven machine learning. *Sci Data*. 2023;10:35. doi:10.1038/s41597-023-01940-7.
19. Colás A, Vigil L, Vargas B, Cuesta-Frau D, Varela M. Detrended fluctuation analysis in the prediction of type 2 diabetes mellitus in patients at risk: model optimization and comparison with other metrics. *PLoS ONE*. 2019;14(12):e0225817. doi:10.1371/journal.pone.0225817.
20. Åm MK, Kölle K, Fougner AL, et al. Effect of sensor location on continuous intraperitoneal glucose sensing in an animal model. *PLoS ONE*. 2018;13(10):e0205447. doi:10.1371/journal.pone.0205447.
21. Tamborlane WV, Beck RW, Bode BW, et al; Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Continuous glucose monitoring and intensive treatment of type 1 diabetes. *N Engl J Med*. 2008;359:1464-1476. doi:10.1056/NEJMoa0805017.
22. Dubosson F, Ranvier JE, Bromuri S, Calbimonte JP, Ruiz J, Schumacher M. The open DINAMO dataset: a multi-modal dataset for research on non-invasive type 1 diabetes management. *Inform Med Unlocked*. 2018;13:92-100. doi:10.1016/j.imu.2018.09.003.
23. Hall H, Perelman D, Breschi A, et al. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biol*. 2018;16:e2005143. doi:10.1371/journal.pbio.2005143.
24. Aleppo G, Ruedy KJ, Riddlesworth TD, et al. REPLACE-BG: a randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes. *Diabetes Care*. 2017;40(4):538-545. doi:10.2337/dc16-2482.
25. Anderson SM, Raghinaru D, Pinsker JE, et al. Multinational home use of closed-loop control is safe and effective. *Diabetes Care*. 2016;39(7):1143-1150. doi:10.2337/dc15-2468.
26. Weinstock RS, DuBose SN, Bergenstal RM, et al. Risk factors associated with severe hypoglycemia in older adults with type 1 diabetes. *Diabetes Care*. 2015;39:603-610. doi:10.2337/dc15-1426.
27. Buckingham B, Beck RW, Tamborlane WV, et al; Diabetes Research in Children Network (DirecNet) Study Group. Continuous glucose monitoring in children with type 1 diabetes. *J Pediatr*. 2007;151:388-393.e2. doi:10.1016/j.jpeds.2007.03.047.
28. The Diabetes Research in Children Network (DirecNet) Study Group. A randomized multicenter trial comparing the gluco-watch biographer with standard glucose monitoring in children with type 1 diabetes. *Diabetes Care*. 2005;28(5):1101-1106. doi:10.2337/diacare.28.5.1101.
29. Tsalikian E, Mauras N, Beck RW, et al. Impact of exercise on overnight glycemic control in children with type 1 diabetes. *J Pediatr*. 2005;147(4):528-534. doi:10.1016/j.jpeds.2005.04.065.
30. Marling C, Bunescu R. The OhioT1DM dataset for blood glucose level prediction. *CEUR Workshop Proc*. 2020;2675:71-74.
31. Public Study Websites. The pediatric artificial pancreas (PEDAP) trial of control-IQ technology in young children in type 1 diabetes. Accessed July 25, 2024. <https://public.jaeb.org/dataset/599>
32. Public Study Websites. Hybrid closed loop therapy and verapamil for beta cell preservation in new onset type 1 diabetes (CLVer). Accessed July 25, 2024. <https://public.jaeb.org/dataset/591>
33. Public Study Websites. A study to assess continuous glucose sensor profiles in healthy non-diabetic participants aged <7 years. Accessed January 4, 2025. <https://public.jaeb.org/dataset/593>
34. Public Study Websites. FLAIR- fuzzy logic automated insulin regulation: a crossover study comparing two automated insulin delivery system algorithms (PID vs. PID + Fuzzy Logic) in individuals with type 1 diabetes. Accessed January 4, 2025. <https://public.jaeb.org/dataset/566>
35. Public Study Websites. The International Diabetes Closed Loop (iDCL) trial: clinical acceptance of the artificial pancreas—a pivotal study of t: slim X2 with control-IQ technology (DCLP3). Accessed January 4, 2025. <https://public.jaeb.org/dataset/573>
36. Public Study Websites. CGM intervention in teens and young adults with T1D (CITY). Accessed January 4, 2025. <https://public.jaeb.org/dataset/565>
37. Freckmann G, Nichols JH, Hinzmann R, et al. Standardization process of continuous glucose monitoring: traceability and

- performance. *Clin Chim Acta*. 2021;515:5-12. doi:10.1016/j.cca.2020.12.025.
38. American Diabetes Association. Understanding diabetes diagnosis. Accessed January 10, 2025. <https://diabetes.org/about-diabetes/diagnosis>
  39. Muniyappa R, Lee S, Chen H, Quon MJ. Current approaches for assessing insulin sensitivity and resistance in vivo: advantages, limitations, and appropriate usage. *Am J Physiol Endocrinol Metab*. 2008;294(1):E15-E26. doi:10.1152/ajpendo.00645.2007.
  40. Godsland IF, Johnston DG, Alberti K, Oliver N. The importance of intravenous glucose tolerance test glucose stimulus for the evaluation of insulin secretion. *Sci Rep*. 2024;14:7451. doi:10.1038/s41598-024-54584-x.
  41. Pacini G, Omar B, Ahrén B. Methods and models for metabolic assessment in mice. *J Diabetes Res*. 2013;2013:986906. doi:10.1155/2013/986906.
  42. Morettini M, Burattini L, Göbl C, Pacini G, Ahrén B, Tura A. Mathematical model of glucagon kinetics for the assessment of insulin-mediated glucagon inhibition during an oral glucose tolerance test. *Front Endocrinol (Lausanne)*. 2021;12:611147. doi:10.3389/fendo.2021.611147.
  43. Piersanti A, Pacini G, Tura A, D'Argenio DZ, Morettini M. An in-silico modeling approach to separate exogenous and endogenous plasma insulin appearance, with application to inhaled insulin. *Sci Rep*. 2024;14:10936. doi:10.1038/s41598-024-61293-y.
  44. Morettini M, Palumbo MC, Göbl C, et al. Mathematical model of insulin kinetics accounting for the amino acids effect during a mixed meal tolerance test. *Front Endocrinol (Lausanne)*. 2022;13:966305. doi:10.3389/fendo.2022.966305.
  45. Subramanian V, Bagger JL, Harihar V, Holst JJ, Knop FK, Villsbøll T. An extended minimal model of OGTT: estimation of  $\alpha$ - and  $\beta$ -cell dysfunction, insulin resistance, and the incretin effect. *Am J Physiol Endocrinol Metab*. 2024;326:E182-E205. doi:10.1152/ajpendo.00278.2023.
  46. De Gaetano A, Panunzi S, Matone A, et al. Routine OGTT: a robust model including incretin effect for precise identification of insulin sensitivity and secretion in a single individual. *PLoS ONE*. 2013;8(8):e70875. doi:10.1371/journal.pone.0070875.
  47. Mari A, Ferrannini E.  $\beta$ -cell function assessment from modelling of oral tests: an effective approach. *Diabetes Obes Metab*. 2008;10(suppl 4):77-87. doi:10.1111/j.1463-1326.2008.00946.x.
  48. Salvatori B, Wegener S, Kotzaeridi G, et al. Identification and validation of gestational diabetes subgroups by data-driven cluster analysis. *Diabetologia*. 2024;67(8):1552-1566. doi:10.1007/s00125-024-06184-7.
  49. Ilari L, Piersanti A, Göbl C, et al. Unraveling the factors determining development of type 2 diabetes in women with a history of gestational diabetes mellitus through machine-learning techniques. *Front Physiol*. 2022;13:789219. doi:10.3389/fphys.2022.789219.
  50. Piersanti A, Giurato F, Göbl C, Burattini L, Tura A, Morettini M. Software packages and tools for the analysis of continuous glucose monitoring data. *Diabetes Technol Ther*. 2023;25(1):69-85. doi:10.1089/dia.2022.0237.
  51. Danne T, Nimri R, Battelino T, et al. International consensus on use of continuous glucose monitoring. *Diabetes Care*. 2017;40:1631-1640. doi:10.2337/dc17-1600.