Published in final edited form as:
Diabetes Res Clin Pract. 2017 December ; 134: 121–130. doi:10.1016/j.diabres.2017.09.009.

Diabetes Alert Dogs (DADs): An Assessment of Accuracy and Implications

Linda A. Gonder-Frederick, Ph.D.[a], **Jesse H. Grabman, B.A.**[a], and **Jaclyn A. Shepard, Psy.D.**[a]

[a]Behavioral Medicine Center, University of Virginia, Charlottesville, Virginia, USA 22908

Abstract

Aims—To test the accuracy of Diabetes Alert Dogs (DADs) by comparing recorded alerts to continuous glucose monitoring (CGM) device readings during waking and sleeping hours.

Methods—14 individuals (7 adults with type 1 diabetes and 7 youth with type 1 diabetes/parents) who owned DADs for ≥ 6 mos wore masked CGM devices over a several-week period while recording DAD alerts electronically and in paper diaries.

Results—During waking hours, sensitivity scores across participants were 35.9% for low BG events and 26.2% for high BG events. DAD accuracy was highly variable with 3/14 individual dogs performing statistically higher than chance. Sensitivity scores were lower during sleep hours of the person with diabetes (22.2% for low BG events and 8.4% for high BG events). DAD accuracy during sleeping hours was also highly variable, with 1/11 individual dogs performing statistically better than chance. Rate of change analyses indicated that DADs were responding to absolute BG level, rather than rapid shifts in glucose levels.

Conclusions—In this study the majority of DADs did not demonstrate accurate detection of low and high BG events. However, performance varied greatly across DADs and additional studies are needed to examine factors contributing to this variability. Additionally, more research is needed to investigate the significant gap between the positive experiences and clinical outcomes reported by DAD owners and the mixed research findings on DAD accuracy.

Keywords

Type 1; Blood Glucose Monitoring; Observational Study; Dogs; Severe Hypoglycemia; CGM

Corresponding Author: Dr. Linda A. Gonder-Frederick, Behavioral Medicine Center, University of Virginia, Box 800223, Charlottesville, Virginia 22908, USA. lag3g@virginia.edu, Phone: +1 (434) 924-5314, Fax: 434-924-0185.

5. Conflict of Interest

5.1. Conflicts of Interest: L.A.G. has previously consulted for Dexcom, Inc. J.H.G. and J.A.S. report no disclosures.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

The importance of blood glucose (BG) monitoring and the detection of extreme BG levels, especially hypoglycemia, for patient safety and quality of life in the management of type 1 diabetes cannot be overestimated. In developed countries, almost all individuals with type 1 diabetes who have sufficient resources/insurance use a BG meter, while others use continuous glucose monitoring (CGM) devices, or both technologies to monitor daily glucose levels.[1] Although invaluable to diabetes self-treatment, both of these devices are invasive, requiring either a finger stick or sensor insertion, and both pose other burdens to the user such as carrying equipment, associated expenses, and calibration. An adjunctive method for BG monitoring is the Diabetes Alert Dog (DAD), which is trained to detect extreme glucose levels, presumably based on olfactory cues the body emits during hypo- and hyperglycemia. The use of DADs has become increasingly popular over the past few years. In fact, a Bing search for the term “Diabetes Alert Dog” returns 5,560 results for the period ranging from 1970 – 2012, but for the four-year period from 2013 to 2016 nearly doubles to 10,800 results. DADs have also received growing attention in the popular media, with almost exclusively and extremely positive testaments from individuals with diabetes and family members regarding the dogs’ abilities and beneficial impact. A recent survey of 135 DAD owners substantiated these positive accounts, with respondents self-reporting numerous benefits including a decrease in diabetes-related hospitalizations, severe hypoglycemic (SH) episodes, and driving accidents after obtaining hypoglycemia detection dogs.[2]

In spite of these positive reports, there is minimal scientific evidence for DAD accuracy and efficacy. Recent studies attempting to test DAD ability to detect extreme BG levels have yielded mixed findings. For example, two studies tested DAD accuracy in highly controlled experimental settings using skin swab samples,[3] or skin swab and breath samples,[4] taken from individuals with type 1 diabetes when BG level was hypoglycemic or euglycemic. The hypo- and euglycemic samples, as well as “blank” samples in one study,[4] were placed in containers in a room where the DAD alert test was conducted with experimenters observing DAD behavior from a separate room. In the first study, a total of 12 hypoglycemic and 12 euglycemic samples (taken from three individuals who were not the DADs’ owners) were presented separately to three trained dogs.[3] Results suggested poor DAD accuracy, with all three dogs showing chance level performance with approximately 50% sensitivity and 50% specificity. In contrast, Hardin et al. tested six DADs over eight trials using a total of 56 samples (8 of which were hypoglycemic), and found higher sensitivity scores ranging from 50.0% to 87.5% (77.6% across dogs), with all DADs performing above statistical chance levels.[4]

Two other studies used observational methods to test DAD performance in real world settings, rather than experimentally controlled situations. One of these analyzed daily diaries in which 18 DAD owners (who had obtained their dog within the past year) recorded daily alerts and BG meter readings.[5] Although Low BG Sensitivity averaged 59.1%, there was a high level of variability across DADs, with accuracy ranging from 33% to 100%. A total of 50%, and 44% of DADs achieved ≥ 65% and 70% Low BG sensitivity, respectively, indicating relatively accurate rates of detection. Another recent study tested accuracy in eight trained dogs by comparing owner-recorded alerts to glucose levels recorded using a

CGM device with readings masked to participants.[6] The study lasted one week and, during this period, collected a total of 45 hypoglycemic events. Although DADs alerted 3 times more often during hypoglycemia as compared to euglycemia, they alerted in a timely manner to only 36% of events, and showed a high rate of false positives. Moreover, CGM alerted prior to DADs in 73% of events verified with a BG meter, leading the authors to conclude that DADs were less useful than CGM devices for hypoglycemia detection.

The purpose of the present study was to conduct a larger trial of DAD accuracy comparing owner-recorded alerts to masked CGM readings in a real world setting over a longer period of time. It was assumed that increasing the number of DADs studied and extending the time period of CGM use to several weeks would increase the number of readings available for analysis, thereby providing a better estimate of DAD accuracy on a day-to-day basis. Additionally, this study is the first to use CGM to investigate DAD accuracy for detection of hyperglycemic BG excursions as well as hypoglycemic, and compare DAD accuracy during the day and night. The primary hypothesis, based on previous results, was that performance would vary greatly across individual DADs, with approximately half demonstrating accuracy above chance levels.

2. Material and methods

2.1. Participants

To control for variation in dog breed and training procedures,[7] the study recruited participants from a single DAD organization. All DADs were Labrador Retrievers bred, raised, and trained in glucose detection for several months by the organization before home placement. Training procedures for BG detection were based on positive reinforcement and utilized standard training stimuli, i.e. blood, perspiration and breath samples from individuals with type 1 diabetes (not the eventual owners) obtained during episodes of hypo- and hyperglycemia. Length of training and DAD age at placement varied. Table 1 shows the age of DADs as well as the number of months each DAD had been placed with its owner. After placement in the home, all owners received continual support and regularly scheduled home training visits by organization staff for a minimum of one year. A total of 8 adult DAD owners with type 1 diabetes and 9 youth with type 1 diabetes/parents participated.

Participants were eligible for the study if they: (1) had type 1 diabetes for at least 1 year and were taking insulin since diagnosis, (2) had a DAD placed in their home for a minimum of 6 months, (3) were not currently using CGM in their diabetes management (so that use of masked CGM would not change routine diabetes care), (4) could complete all study protocol tasks, including filling out questionnaires and diaries in English, (5) had access to an internet-connected computer compatible with study software, (6) ranged in age between 6–65 years, and (7) were willing to avoid consumption of acetaminophen-containing products for the duration of the study to maintain accuracy of sensor readings as per manufacturer instructions. Exclusion criteria included pregnancy and a history of deep tissue infection. After the consent/assent process, it was discovered that two youth/parents did not meet inclusion criteria, and these families' data were excluded from analysis. One adult participant dropped out of the study before initiating CGM data collection. The final sample consisted of 7 adults (adult age median = 34.0 yrs [range = 22–43 yrs]; 6 female; median self-reported HbA1c = 7.2%, or 51 mmol/mol [range = 5.9 – 8.9%, or 41–74 mmol/mol]; diabetes duration median = 22 yrs [range = 9 – 35 yrs]) and 7 youth/parents (youth age median = 10.0 [range = 8–17 yrs]; 3 female; Median self-reported HbA1c = 8.1%, or 64 mmol/mol, [range = 6.8 – 9.0%, or 51–75 mmol/mol]; diabetes duration median = 5 yrs [range = 3–8 yrs]). All participants reported current use of an insulin pump. All adults and the majority of youth/parents (5/7) reported previous use and discontinuation of CGM. DADs ranged in age from 14 – 104 months (Median = 28) and were placed in participants' homes for 6–41 months (Median = 20) prior to the study. Participants were highly confident in their DADs, with a median perceived accuracy of 80% (Range 50–90%), and most (13/14) indicating that their DAD was 'often' or 'always' more accurate than current diabetes technology. The study protocol was approved by the University of Virginia Institutional Review Board for Health Sciences Research.*2.2. Procedure*

The training organization initially approached adult and parent DAD owners who met inclusion/exclusion criteria to gauge interest in study participation, and then provided contact information to the study team for those who were interested. A study team member then contacted potential participants to provide information about the project. Eligible individuals/families attended an orientation session in cohorts of 1–4 DAD owners.

At the orientation, DAD owners reviewed study aims and procedures, and had the opportunity to ask questions before signing informed consent (Adults and Primary caregiver) or assent (youth) forms. Participants were asked to use a masked CGM device (Dexcom G4; San Diego, CA) for 4 weeks while maintaining their normal daily diabetes care, calibrate the device according to manufacturer specifications, and replace the sensor weekly. Training for use of the Dexcom G4 was accomplished by watching an instructional video and verbally reviewing the instructions, followed by insertion of a sensor and beginning the initial phase of calibration during the meeting. Participants were also provided with the phone number of the study team and device manufacturer for technical support. Once comfortable with operating the device, participants (and/or parents) entered DAD alerts electronically using the CGM's event recorder function, and also completed daily paper diaries to provide more detailed data about DAD alerts (e.g. specific alert behaviors). DAD alert behaviors varied by owner; however common indicators reported by participants included barking and/or pawing/nudging the person with diabetes. Additionally, participants recorded time ranges for when they were not in the same location as their DAD (e.g. when at work/school without their DAD), as well as hours when the person with diabetes was asleep. Individuals with diabetes wore CGM devices for up to 6 weeks (in cases of sensor failure), and participants were reminded weekly to send electronic CGM data and diary scans to a secure email address. Participants received up to $200.00 in remuneration for their time and effort: $40.00/wk (up to 4 weeks) for CGM use, and $40.00 for completed post-questionnaires.

2.3. Data Analysis

While CGM offers unrivaled access to participant glucose dynamics, this approach presents difficulties when assessing DAD accuracy. First, DAD behaviors are a dichotomous variable (Alert/ No Alert), whereas CGM values are an interval variable ranging from 40 – 400 mg/dl (2.2 – 22.2 mmol/L). To address this, we used Signal Detection Theory[8,9] methods to classify accuracy into four categories based on participant BG value (within target vs. out of range) and whether or not the DAD alerted:

1. *Hits* (BG Low/High; DAD alerted). Primary analyses examined BG thresholds ≤ 70 mg/dl (3.9 mmol/L; hypoglycemia) and ≥ 180 mg/dl (10.0 mmol/L; hyperglycemia), as suggested by current clinical guidelines.[10] As supplementary analyses, we assessed DAD accuracy in response to more extreme hypo- (≤ 54 mg/dl; ≤ 3.0 mmol/L) /hyperglycemia (≥ 250 and 300 mg/dl; ≥ 13.9 and 16.7 mmol/L).
2. *Misses* (BG Low/High; no DAD alert)
3. *Correct Rejections* (BG in target range; no DAD alert). We defined target BG as ranging from 70 - 180 mg/dl (3.9 – 10.0 mmol/L).
4. *False Alarms* (BG in target range; DAD alert).

Another issue to resolve before computing DAD accuracy statistics is that DAD alerts are relatively infrequent compared to CGM readings (occurring every 5 mins), which in turn could lead to overestimates in *Misses* and *Correct Rejections*. For example, a DAD may alert once in 30 mins of hypoglycemia, but CGM generates six readings. Without adjustment, this results in 1 *Hit* and 5 *Misses* – significantly underestimating DAD accuracy.

To address this issue, we defined an *event* as a period of low/target/high BG lasting ≥ 15 minutes, when the owner indicated they were in the same location as their DAD. Thus, in the example above, the DAD receives credit for one *Hit* for the entire 30-min hypoglycemic event. Additionally, we imposed the constraint that alerts occur within 20 minutes before (or after) the first out of range BG reading to qualify as a Hit.. We implemented this criterion to account for alert clinical utility (e.g. ability for the person with diabetes to self-treat before experiencing severe mental disorientation or unconsciousness), time lag of CGM readings,[11] and owner reports that DADs often alert ahead of BG extremes. We analyzed events during waking and sleeping hours separately. For three participants who did not provide sleep data, a conservative period ranging from 8:00 AM – 8:00 PM was used to approximate waking hours and their nighttime hours were excluded from data analysis.

After categorizing all BG events, we calculated the following primary outcome measures to summarize DAD performance:

1. *Overall Sensitivity* is the percentage of out-of-range BG events with a DAD alert (*Hits*) compared to the total number of out-of-range events (*Hits* + *Misses*). We separately evaluated low and high BG sensitivity.
2. *Specificity* is the percentage of target BG events with no DAD alert (*Correct Rejections*) compared to the total number of target events (*Correct Rejections* + *False Alarms*).
3. *Positive Predictive Value (PPV)* is the percentage of accurate alert events (*Hits*) compared to the total number of DAD alert events (*Hits + False Alarms*).

4. *Overall accuracy* is the percentage of accurate events (*Hits* + *Correct Rejections*) compared to the total number of events (*Hits + Correct Rejections + False Alarms + Misses)*.

5. *d'* quantifies the dual aims of remaining sensitive (alerting to low/high BG), while also remaining specific (not falsely alerting).[12] Values > 0.00 indicate above chance accuracy, with higher (positive) values corresponding to better overall performance.

3. Results

3.1. CGM Data

Participants wore the CGM between 13 – 50 days (Median = 29) and obtained 3007 – 11639 CGM readings (Median = 7430) (see Supplemental Table S1 for individual participant results). Adherence to CGM use was good, with the person with diabetes wearing the device approximately 78.9% – 94.5% (Median = 88.4%) of the time. Overall, participants spent a slight majority of the time in target range (Median = 50.7%; Range = 20.7 – 66.9%), followed by hyper- (Median = 45.0%; Range = 13.6 – 78.8%), and hypoglycemia (Median = 4.3%; Range = 0.2 – 20.7%). We found no substantive differences in accuracy for more extreme BG thresholds (e.g. ≤ 54 mg/dl; ≤3.0 mmol/L), both during waking and sleeping hours, therefore we only present results for the range between 70 – 180 mg/dl (3.9 – 10.0 mmol/L).

Waking Hour Accuracy—Supplemental Table 1 shows the frequencies and proportions of events and accuracy categories for each participant. When awake, the total number of analyzable hypoglycemic events per participant ranged between 0.48 – 21.50/week (Mean = 5.73, Median = 3.08). The frequency of hyperglycemic events ranged from 6.25–18.31/week (Mean = 13.12, Median = 13.06). “Target” events were most common (51.3% of all events; Range 40.6 – 59.3%), with fewer hyper- (33.9% of all events; Range = 14.9 – 54.9%) and hypoglycemic (14.8% of all events; Range = 1.4 – 33.0%) events.

Table 1 presents accuracy measures for each DAD. Overall sensitivity across DADs was 29.1%, with improved performance when alerting to low BGs (35.9%) compared to high BGs (26.2%). Overall, DADs did better avoiding unnecessary alerts, as reflected by higher scores on specificity (65.5%) and PPV (61.5%). However, overall accuracy was low (47.8%), with only 3/14 DADs performing at levels above statistical chance (*d'* = −.150 across dogs).

Accuracy was highly variable across DADs. Overall sensitivity ranged from 6.7% – 45.6%, low BG sensitivity from 0.0 – 100.0%, and high BG sensitivity from 5.9 – 46.2%. Specificity ranged from 43.4 – 87.2% and PPV from 32.6 – 77.8%. Overall accuracy ranged from 39.1 – 57.6%.

3.2. Accuracy when Participants were Asleep

Supplemental Table S2 presents frequencies and proportions of events and accuracy categorizations for when participants indicated that they were asleep. The total number of nocturnal hypoglycemic events was substantially lower compared to diurnal events (*M* = 5.38/week), ranging from 0.00 – 2.41/week (Mean = .65, Median = .50). Hyperglycemic events ranged from 0.25 – 9.25/week (Mean = 4.06, Median = 6.52). Compared to waking hours, participants had a greater proportion of nocturnal hyperglycemic events (43.8% of total events; Range = 10.0 – 64.3%), with diminished percentages of Target (49.3% of total events; Range = 35.7 – 70.0%) and hypoglycemic (7.0% of total events; Range = 0.0 – 20.0%) events.

Table 2 displays each DAD's nocturnal accuracy measures. DADs showed minimal sensitivity for both low (22.2%) and high (8.4%) BG, resulting in an overall sensitivity of 10.9%. Compared to waking hours there was a lower rate of DAD false alarms, with increased specificity (83.0%) and PPV (68.9%) values, which may reflect either the DAD or owner sleeping. Overall accuracy remained similar to waking hours (46.6%), with *d'* values supporting statistically above chance accuracy for 1/11 DADs (*d'* overall = −.281).

As in the waking hours analysis, DADs varied in their nocturnal detection abilities. Overall sensitivity ranged from 0.0 – 22.2%, hypoglycemic sensitivity from 0.0 – 50.0%, and hyperglycemic sensitivity from 0.0 – 22.2%. Specificity ranged from 40.0 – 96.7% and PPV from 33.3 – 80.0%. Overall accuracy ranged from 28.6 – 55.1%.

3.3. An alternative Event Blocking Scheme

Due to concerns that the event blocking scheme could underestimate accuracy by including events of hypoglycemia that were too transient for the DAD to be given sufficient time to detect them (or were the result of CGM artifacting), we performed an additional analysis on waking hours that adjusted the block period to 45 minutes from the original 15 minutes. DADs were given credit for alerts 30 minutes before to 45 minutes after the first low/high BG event. The more liberal time criteria for accurate alerting resulted in an increase in overall accuracy to 39.4% and hypoglycemia sensitivity to 46.3% (data not shown), though it should be noted that 3 participants no longer reported any hypoglycemic events due to the stricter blocking criteria. However, we only found evidence for one additional DAD achieving statistical accuracy using the d' statistic, indicating that event blocking did not fully explain DAD inaccuracy.

3.4. DAD Detection of Rate of Change vs. Current BG Level

To assess whether DAD alerts correspond to a rapid rate of change in BG (RoC), rather than raw BG value, we computed RoC for the 1,919 DAD alerts that met analysis criteria as defined by Clarke and Kovatchev (2009).[13] Figure 1 shows the frequency distribution of RoC scores, ranging from −5 to +5 mg/dl/min (− .28 to + .28 mmol/L/min). Overall, DADs responded to a mean absolute value RoC of ± .98 mg/dl/min (SD = 1.00) (± .05 mmol/L/min, SD = .06). The majority of alerts (66.0%) occurred when BG was changing by < 1 mg/dl/min (< .06 mmol/L/min). Fewer alerts occurred when BG rate of change was 1–2 (22.5%), 2–3 (7.0%), and >3 (4.6%) mg/dl/min (.06 – .11, .11–.17, and > .17 mmol/L/min). When BG was Low, DAD alerts corresponded to a slight downward BG trajectory (*M* = − .47 mg/dl/min, SD = .88; – .03 mmol/L/min, SD = .05), while alerts to hyperglycemia generally occurred during slight BG increases (*M* = + .27 mg/dl/min, SD = 1.55; + .01 mmol/L/min, SD = .09).

As a comparative analysis (and as done by Los et al., 2016),[6] we graphed the proportion of DAD alerts at each CGM value, and contrasted this with proportions of all CGM readings at these values (Figure 2). As indicated by higher bars on the graph, DADs alerted more frequently during hypo- and hyperglycemia than would be expected by participants' total proportion of readings in these ranges. Given that BG did not usually fluctuate rapidly during DAD alerts, and high Positive Predictive/Specificity values, it appears that DADs respond to current BG levels rather than RoC.

4. Discussion

This study did not find evidence to support the hypothesis that DADs accurately detect extreme BG levels, either hypo- or hyperglycemic. In fact, our results for hypoglycemia sensitivity during waking hours (35.9% across DADs) were near identical to those reported by Los et al. in their study comparing DAD alerts to CGM data (36%).[6] Hyperglycemia sensitivity scores for waking hours were lower than those for hypoglycemia. Although these results replicated our previous finding that DAD accuracy is highly variable across individual dogs,[5] in this study fewer DADs (3/14) performed at levels statistically higher than chance during waking hours. Sensitivity scores were lower during sleeping hours for both hypo- and hyperglycemia and, although DAD accuracy was also variable under these conditions, only one dog performed statistically higher than chance. DAD accuracy during the night is particularly important from a clinical perspective since nocturnal hypoglycemia is highly prevalent and of great concern to many patients. It is possible that these results may reflect owners sleeping through DAD alerts, or DADs sleeping through nocturnal hypoglycemic episodes. However, it is also the case that several participants did not experience nocturnal hypoglycemia during the study which may have contributed to lower estimates of sensitivity during the night. Nonetheless, the overall findings from this study do not support the belief that the accuracy and reliability of DAD performance is comparable to technological devices used for BG monitoring.

Clearly there is a sizeable gap between the picture of DAD accuracy that has emerged from the majority of recent scientific studies compared to the reports of people with diabetes who own and use DADs to assist with BG monitoring on a daily basis. Only one recent study, using a highly controlled experimental procedure with skin swab and breath samples showed high levels of DAD accuracy.[4] However, in spite of the weak scientific evidence for DAD accuracy, it is premature to simply discount the positive experiences reported by so many individuals and families who use DADs. Possible contributing factors to this discrepancy need to be considered carefully, including methodological limitations inherent in both experimental and observational approaches to the study of DAD accuracy. For example, the experimental approach allows the control of potentially confounding factors such as differences in the testing environment or target stimuli to be detected. However, the relevance of findings obtained in this controlled environment to DAD accuracy in natural settings with their owner can be questioned.

An obvious potential problem in our study procedure is the reliance on owner self-report data. For example, it is highly likely that study participants did not record every alert that occurred over the course of the study, which could result in underestimates of DAD accuracy. A post hoc analysis of our data found a a positive correlation between DAD Sensitivity scores and the average number of daily diary entries recorded by owners, providing some support for this hypothesis (r_τ = .42, p = .019, one-tailed). In addition, although participants recorded periods where the person with diabetes was away from their DAD, this study could not objectively determine actual physical proximity (either during waking or sleeping hours), which potentially diminishes sensitivity measures by overestimating misses. Moreover, in spite of this study not finding a correlation between the percentage of time participants reported to be with the DAD and waking hypo-/hyperglycemia sensitivity (r_τ = .15, p = .246 and r_τ = .00, p = .500, one-tailed ,respectively), it is possible that DADs with more daily exposure to the person with diabetes exhibit better accuracy through increased opportunity for training/reinforcement.

Compounding the issues with self-report data, the strict criteria adopted for timing of DAD alerts (e.g. 20 min after the first CGM low BG reading) may also contribute to accuracy underestimation. A post hoc analysis to test this hypothesis showed that waking hour sensitivity scores increased to 56.1% and 58.8% for hypo- and hyperglycemia, respectively, when DAD alerts at any time during an event were counted. However, the clinical utility of alerts occurring 30 – 60 minutes after the onset of a hypo-/hyperglycemic episode is debatable. Finally, although participants in this study wore CGM for the longest period to date for assessing DAD performance, some exhibited a low number of BG events (i.e. only 1 or 2 hypoglycemic episodes), which could have contributed to under/over-estimations of individual DAD accuracy.

In addition to methodological factors, there are also well-documented psychological processes that could potentially bias DAD owner perceptions, resulting in overestimations of accuracy. These include confirmatory cognitive biases (e.g. selective recall for instances when DADs are accurate),[14] or cognitive dissonance (e.g. biases due to level of investment in DAD accuracy).[15] Exceptionally memorable cases (such as DADs alerting prior to a traumatic severe hypoglycemic episode) could amplify these effects. However, it seems unlikely that such biases would completely account for the large discrepancy between scientific findings and owners' experiences. In addition, it is important to remember that individual DADs appear to vary greatly in accuracy, which could also contribute to mixed scientific findings and discrepancies. It is critical to investigate the factors contributing to these individual differences in DAD performance. For example, do these differences arise secondary to differences in dogs' inherent abilities to detect glycemic changes in humans based on olfactory cues, or perhaps from differences in the training process? While this study attempted to control for variations in training procedures by testing DADs from one organization, it will be necessary in the future to investigate the impact of different training techniques, e.g. length and type of training and criteria for deciding that DADs have achieved adequate accuracy.

In addition, what role, if any, do differences in the skill of the DAD owner/handler play in the dog's performance after placement? For example, owner skill in accurate recognition of

DAD alerts or delivery of timely and appropriate reinforcement? A post-hoc analysis of our data revealed that participants who owned their DADs for longer periods of time showed higher overall sensitivity (r_τ = 0.42, p = .018, one-tailed), with diminished specificity (r_τ = -0.48, p = .009, one-tailed). This suggests an increase over time in alert behaviors possibly due to DADs receiving more frequent reinforcement for alert behaviors irrespective of the accuracy of the alerts. However, more research is necessary to explore this issue.

Clearly, DAD performance could be affected by a number of complex and interacting factors that need to be better understood. However, it is important to note that previous findings also point to benefits in clinical outcomes with DAD use, such as fewer self-reported diabetic ketoacidosis hospitalizations and SH episodes.[2] Given that high levels of accuracy are so far unsubstantiated in research, what could account for these improvements? One possibility is that DADs serve as a tool for increasing engagement in diabetes self-management and awareness. For example, the DAD's presence throughout the day may remind owners to engage in more frequent BG checks, which have historically been linked to improved glycemic outcomes.[16,17] These potential benefits should not be readily discounted and deserve more research.

In summary, this comparison of DAD alerts to CGM data collected by 14 individual DAD owners over a several-week period did not find support for high levels of accuracy. Much more research is needed that carefully considers the complexity of DAD training, performance and interaction with owners, as well as the methodological issues that can affect results. Larger studies that can assess potentially relevant DAD and owner attributes may greatly expand our understanding of variability in accuracy. In addition, use of alternative methodological approaches is also needed, including the use of video recording to allow the direct observation of DAD-owner interactions. Organizations that train and provide DADs for individuals with diabetes need to work closely with the scientific community to conduct this research. In the meantime, these organizations should consider the development of regulatory standards for DAD training and criteria for adequate performance before placement, as well as requirements to assess performance after placement. One way to assess DAD performance is comparison of dog alerts with CGM measures which may be a useful criterion from a regulatory perspective. Finally, it is important that DAD owners follow guidelines used with earlier CGM devices, which were not approved for clinical decision making until recently,[18] and confirm alerts with measurement of BG levels prior to any self-treatment behaviors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

6.1. Author Contributions: L.A.G. conceived and designed the project, served as the study principal investigator, and is the guarantor of the work; additionally conducted participant orientation meetings, supervised data collection and analysis, participated in data interpretation, and wrote sections of the manuscript. JHG conducted participant orientation meetings, assisted in the implementation of methodological design and procedures, performed data management and analysis, participated in data interpretation, and wrote sections of the manuscript. JAS prepared

study materials, assisted in the implementation of methodological design and procedures, participated in data interpretation, and reviewed/edited the manuscript.

6.2. Acknowledgements: The authors wish to acknowledge Dallas Ducar, B.A., Zachary McElgunn, B.A., Joseph E. Poler, Jr., Psy.D., and Anand Tripathi, B.A. for assisting with participanat orientation meetings, data coding, and for providing thoughtful discussion on the project. Additionally, we thank Matthew Moncrief, B.S. for assistance with data analysis. Finally, the authors thank Barbara Anderson, Ph.D. and Christel Hendrieckx, Ph.D. for their helpful and thought provoking comments on a preliminary version of the manuscript.

6.3. Funding sources: This work was supported by NIH-NIDDK [Grant #1R21DK099697-01].

References

1. Gonder-Frederick LA, Shepard JA, Grabman JH, Ritterband LM. Psychology, technology, and diabetes management. Am Psychol. 2016; 71(7):577. [PubMed: 27690486]
2. Petry NM, Wagner JA, Rash CJ, Hood KK. Perceptions about professionally and non-professionally trained hypoglycemia detection dogs. Diabetes Res Clin Pract. 2015; 109(2):389–96. [PubMed: 26044610]
3. Dehlinger K, Tarnowski K, House JL, Los E, Hanavan K, Bustamante B, Ahmann AJ, Ward WK. Can trained dogs detect a hypoglycemic scent in patients with type 1 diabetes? Diabetes Care. 2013; 36(7):e98–9. [PubMed: 23801820]
4. Hardin DS, Anderson W, Cattet J. Dogs can be successfully trained to alert to hypoglycemia samples from patients with type 1 diabetes. Diabetes Ther. 2015; 6(4):509–17. [PubMed: 26440208]
5. Gonder-Frederick LA, Grabman JH, Shepard JA, Tripathi AV, Ducar DM, McElgunn ZR. Variability of Diabetes Alert Dog Accuracy in a Real-World Setting. J Diabetes Sci Technol. 20171932296816685580
6. Los EA, Ramsey KL, Guttmann-Bauman I, Ahmann AJ. Reliability of Trained Dogs to Alert to Hypoglycemia in Patients With Type 1 Diabetes. J Diabetes Sci Technol. 20161932296816666537
7. Gonder-Frederick LA, Ducar D, Grabman JH, Shepard J. Diabetes alert dogs: A review of the industry [Abstract]. Diabetes. 2014; 63(suppl):A223.
8. Swets, J. Signal detection and recognition by human observers. New York: Wiley; 1964.
9. Swets, J. Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. Hillsdale, N.J.; England: Lawrence Erlbaum Associates, Inc; 1996.
10. American Diabetes Association. Standards of Medical Care in Diabetes 2017. Diabetes Care. 2017; 40(Supplement 1):S48–56. [PubMed: 27979893]
11. Schmelzeisen-Redeker G, Schoemaker M, Kirchsteiger H, Freckmann G, Heinemann L, del Re L. Time delay of CGM sensors: relevance, causes, and countermeasures. J Diabetes Sci Technol. 2015; 9(5):1006–15. [PubMed: 26243773]
12. Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behav Res Methods Instrum Comput. 1999; 31(1):137–49. [PubMed: 10495845]
13. Clarke W, Kovatchev B. Statistical tools to analyze continuous glucose monitor data. Diabetes Technol The. 2009; 11(suppl 1):S45–S54.
14. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. Rev Gen Psychol. 1998; 2(2):175.
15. Harmon-Jones E, Harmon-Jones C. Cognitive dissonance theory after 50 years of development. Soc Psychol (Gott). 2007; 38(1):7–16.
16. Miller KM, Beck RW, Bergenstal RM, Goland RS, Haller MJ, McGill JB, Rodriguez H, Simmons JH, Hirsch IB. T1D Exchange Clinic Network. Evidence of a strong association between frequency of self-monitoring of blood glucose and hemoglobin A1c levels in T1D exchange clinic registry participants. Diabetes Care. 2013; 36(7):2009–14. [PubMed: 23378621]
17. Schütt M, Kern W, Krause U, Busch P, Dapp A, Grziwotz R, Mayer I, Rosenbauer J, Wagner C, Zimmermann A, Kerner W. Is the frequency of self-monitoring of blood glucose related to long-term metabolic control? Multicenter analysis including 24500 patients from 191 centers in Germany and Austria. Exp Clin Endocrinol Diabetes. 2006; 114(07):384–8. [PubMed: 16915542]

18. US Food and Drug Administration. Approval Order: Dexcom G5 Mobile Continuous Glucose Monitoring System. PI20005. Department of Health and Human Services; Silver Spring, MD: 2016.

Highlights

- Largest and most comprehensive study to date investigating real-world Diabetes Alert Dog (DAD) accuracy using continuous glucose monitoring.
- First study to systematically assess DAD accuracy of hypo-/hyperglycemia detection during the daytime and nighttime.
- Discusses substantive methodological issues and implications relevant to DAD research and use.

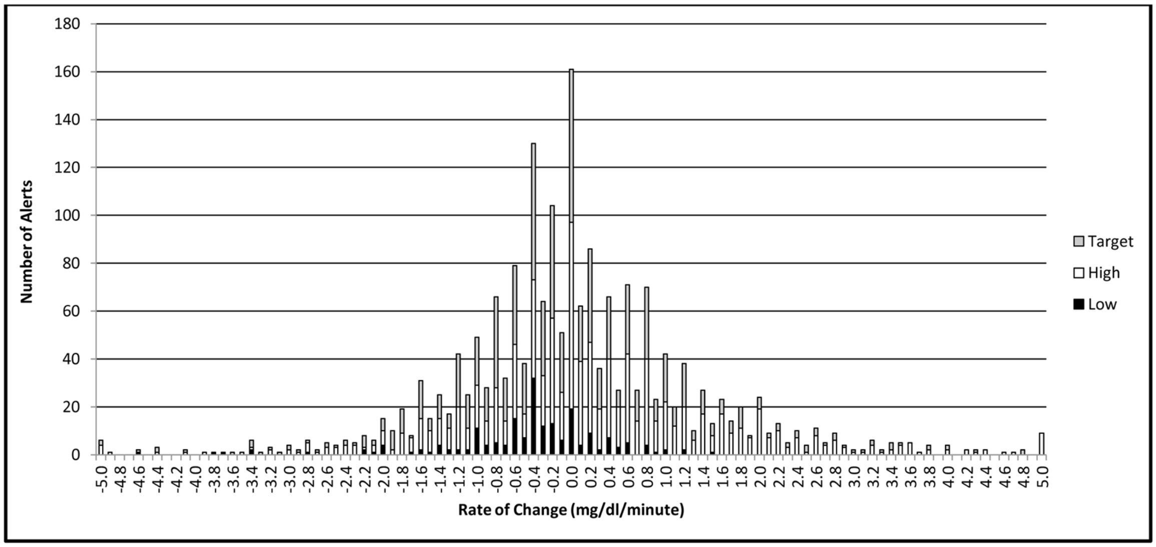

Figure 1.

Frequency distribution of BG rate of change at time of DAD alert.

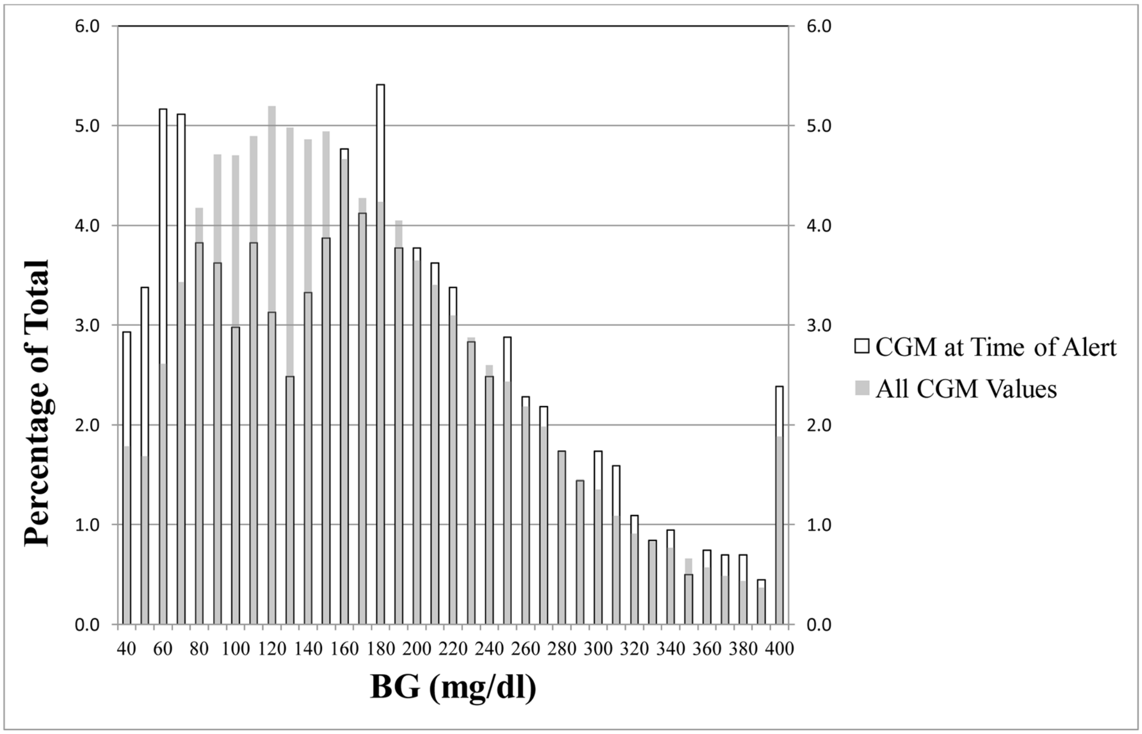

Figure 2.

Proportion of DAD alerts at each CGM value compared to proportions of all CGM readings at these values.

Table 1

Accuracy results when the participant was awake.

Age Group	DAD Age (mo)	DAD Time with Owner (mo)	Total Events	Hits	Misses	False Alarms	Correct Rejections	Low Sensitivity	High Sensitivity	Overall Sensitivity	Specificity	PPR	*d’*
Adult	23	8	74	3 (4.1%)	32 (43.2%)	5 (6.8%)	34 (45.9%)	100.0%	5.9%	8.6%	87.2%	70.6%	−0.232
Adult	14	6	109	10 (9.2%)	44 (40.4%)	16 (14.7%)	39 (35.8%)	18.2%	18.8%	18.5%	70.9%	66.0%	−0.344
Adult	28	22	97	12 (12.4%)	29 (29.9%)	13 (13.4%)	43 (44.3%)	43.8%	20.0%	29.3%	76.8%	65.8%	0.187
Adult	29	26	261	57 (21.8%)	68 (26.1%)	77 (29.5%)	59 (22.6%)	45.3%	46.2%	45.6%	43.4%	54.4%	−0.277
Adult	43	40	167	35 (21.0%)	43 (25.7%)	43 (25.7%)	46 (27.5%)	75.0%	43.2%	44.9%	51.7%	58.7%	−0.086
Adult	21	13	270	43 (15.9%)	90 (33.3%)	46 (17.0%)	91 (33.7%)	50.0%	25.3%	32.3%	66.4%	66.9%	−0.034
Adult	28	23	389	49 (12.6%)	145 (37.3%)	60 (15.4%)	135 (34.7%)	28.6%	21.9%	25.3%	69.2%	58.6%	−0.163
Youth	104	31	128	24 (18.8%)	34 (26.6%)	33 (25.8%)	37 (28.9%)	25.0%	44.0%	41.4%	52.9%	54.2%	−0.146
Youth	99	18	135	9 (6.7%)	46 (34.1%)	31 (23.0%)	49 (36.3%)	20.0%	15.0%	16.4%	61.3%	32.6%	−0.693
Youth	17	6	83	3 (3.6%)	42 (50.6%)	5 (6.0%)	33 (39.8%)	0.0%	7.0%	6.7%	86.8%	75.0%	−0.381
Youth	15	10	128	12 (9.4%)	51 (39.8%)	14 (10.9%)	51 (39.8%)	30.0%	17.0%	19.0%	78.5%	65.9%	−0.088
Youth	17	8	139	24 (17.3%)	41 (29.5%)	18 (12.9%)	56 (40.3%)	33.3%	37.3%	36.9%	75.7%	71.4%	0.363
Youth	43	41	133	28 (21.1%)	51 (38.3%)	30 (22.6%)	24 (18.0%)	50.0%	34.2%	35.4%	44.4%	67.0%	−0.513
Youth	34	30	165	14 (8.5%)	70 (42.4%)	12 (7.3%)	69 (41.8%)	28.0%	11.9%	16.7%	85.2%	77.8%	0.078
Total			2278	14.2%	34.5%	17.7%	33.6%	35.9%	26.2%	29.1%	65.5%	61.5%	−0.150
Median			134	12.5%	35.7%	15.1%	36.0%	31.7%	20.9%	27.3%	70.1%	65.9%	−0.155
Minimum			74	3.6%	25.7%	6.0%	18.0%	0.0%	5.9%	6.7%	43.4%	32.6%	−0.693
Maximum			389	21.8%	50.6%	29.5%	45.9%	100.0%	46.2%	45.6%	87.2%	77.8%	0.363

Table 2

Accuracy results when the participant was asleep.

Age Group	DAD Age (mo)	DAD Time with Owner (mo)	Total Events	Hits	Misses	False Alarms	Correct Rejections	Low Sensitivity	High Sensitivity	Overall Sensitivity	Specificity	PPR	*d’*
Adult	23	8	22	0 (.0%)	9 (40.9%)	2 (9.1%)	11 (50.0%)	-	0.00%	0.00%	84.60%	33.30%	−2.699
***Adult**	14	6	-	-	-	-	-	-	-	-	-	-	-
Adult	28	22	49	0 (.0%)	21 (42.9%)	1 (2.0%)	27 (55.1%)	0.00%	0.00%	0.00%	96.40%	50.00%	−1.915
Adult	29	26	10	0 (.0%)	3 (30.0%)	2 (20.0%)	5 (50.0%)	0.00%	0.00%	0.00%	71.40%	33.30%	−3.153
Adult	43	40	77	5 (6.5%)	32 (41.6%)	6 (7.8%)	34 (44.2%)	0.00%	13.90%	13.50%	85.00%	76.00%	−0.065
***Adult**	21	13	-	-	-	-	-	-	-	-	-	-	-
Adult	28	23	23	0 (.0%)	13 (56.5%)	1 (4.3%)	9 (39.1%)	0.00%	0.00%	0.00%	90.00%	80.00%	−2.437
Youth	104	31	73	2 (2.7%)	36 (49.3%)	6 (8.2%)	29 (39.7%)	0.00%	5.40%	5.30%	82.90%	45.50%	−0.67
Youth	99	18	76	5 (6.6%)	33 (43.4%)	5 (6.6%)	33 (43.4%)	37.50%	6.70%	13.20%	86.80%	78.30%	0.001
Youth	17	6	64	1 (1.6%)	34 (53.1%)	1 (1.6%)	28 (43.8%)	0.00%	3.20%	2.90%	96.60%	80.00%	−0.081
Youth	15	10	67	0 (.0%)	37 (55.2%)	1 (1.5%)	29 (43.3%)	-	0.00%	0.00%	96.70%	80.00%	−1.884
Youth	17	8	71	6 (8.5%)	31 (43.7%)	8 (11.3%)	26 (36.6%)	50.00%	12.10%	16.20%	76.50%	74.20%	−0.263
Youth	43	41	14	2 (14.3%)	7 (50.0%)	3 (21.4%)	2 (14.3%)	-	22.20%	22.20%	40.00%	72.70%	−1.017
***Youth**	34	30	-	-	-	-	-	-	-	-	-	-	-
Total			**546**	**3.8%**	**46.9%**	**6.6%**	**42.7%**	**22.2%**	**8.4%**	**10.9%**	**83.0%**	**68.9%**	**−0.281**
Median			**64.0**	**1.6%**	**43.7%**	**7.8%**	**43.4%**	**0.0%**	**3.2%**	**2.9%**	**85.0%**	**74.2%**	**−1.02**
Minimum			**10**	**0.0%**	**30.0%**	**1.5%**	**14.3%**	**0.0%**	**0.0%**	**0.0%**	**40.0%**	**33.3%**	**−3.153**
Maximum			**77**	**14.3%**	**56.5%**	**21.4%**	**55.1%**	**50.0%**	**22.2%**	**22.2%**	**96.7%**	**80.0%**	**0.001**

* Participant Did not complete sleep diary.