Journal of Biomedical Informatics 63 (2016) 259–268

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using uncertain data from body-worn sensors to gain insight into type 1 diabetes

Nathaniel Heintzman [a], Samantha Kleinberg [b,*]

[a] Dexcom Inc., San Diego, CA, United States
[b] Department of Computer Science, Stevens Institute of Technology, NJ, United States

ARTICLE INFO

Article history:
Received 10 March 2016
Revised 19 August 2016
Accepted 26 August 2016
Available online 28 August 2016

Keywords:
Diabetes Mellitus, Type 1
Monitoring, Ambulatory
Uncertainty
Causality

ABSTRACT

The amount of observational data available for research is growing rapidly with the rise of electronic health records and patient-generated data. However, these data bring new challenges, as data collected outside controlled environments and generated for purposes other than research may be error-prone, biased, or systematically missing. Analysis of these data requires methods that are robust to such challenges, yet methods for causal inference currently only handle uncertainty at the level of causal relationships – rather than variables or specific observations. In contrast, we develop a new approach for causal inference from time series data that allows uncertainty at the level of individual data points, so that inferences depend more strongly on variables and individual observations that are more certain. In the limit, a completely uncertain variable will be treated as if it were not measured. Using simulated data we demonstrate that the approach is more accurate than the state of the art, making substantially fewer false discoveries. Finally, we apply the method to a unique set of data collected from 17 individuals with type 1 diabetes mellitus (T1DM) in free-living conditions over 72 h where glucose levels, insulin dosing, physical activity and sleep are measured using body-worn sensors. These data often have high rates of error that vary across time, but we are able to uncover the relationships such as that between anaerobic activity and hyperglycemia. Ultimately, better modeling of uncertainty may enable better translation of methods to free-living conditions, as well as better use of noisy and uncertain EHR data.

1. Introduction

Chronic diseases like diabetes, hypertension, and heart disease account for most U.S. deaths each year [1], and unlike acute illnesses they are primarily managed not by clinicians but by patients themselves. Blood glucose management in people with type 1 diabetes mellitus (T1DM) imposes a particularly significant cost on time and attention. Patients make frequent (∼hourly) decisions about the timing and dosing of insulin to manage their blood glucose with infrequent (∼semiannually) feedback from clinicians while having to account for many factors that affect glucose including physical activity, meals, and illness. Managing blood glucose is key to avoiding secondary complications like stroke, but requires constant vigilance from patients with little external feedback and support.

Yet, patients now generate massive amounts of health data through tracking symptoms and disease progression, uploading sensor data into patient portals, and sharing information with one another through social networks like PatientsLikeMe [2]. These patient-generated data may fill in information gaps between medical visits and help engage patients but there have been concerns about data quality and how to integrate patient-generated data into care without overloading clinicians [3]. Tests with providers, though, found the primary question was whether the data provided actionable information [4].

Most work has focused on how these data can improve encounters with providers, but continuously-collected data from body-worn sensors and devices such as continuous glucose monitors (CGMs), insulin pumps, and activity monitors could be used as input to a patient-centered decision support system, which would reduce the burden placed on patients with diabetes and other chronic diseases. While CGMs allow continuous recording of glucose data, these and other sensors can have errors, and imperfect use in real-world settings leads to new challenges (e.g. samples contaminated by inadequate hand-washing, loss of connectivity). Further, not all sensors can be worn during all activities (e.g. swimming) or may be removed by a patient for other reasons, leading to missing data. These issues are present in many types of

* Corresponding author.
E-mail addresses: nheintzman@dexcom.com (N. Heintzman), samantha.kleinberg@stevens.edu (S. Kleinberg).

http://dx.doi.org/10.1016/j.jbi.2016.08.022
1532-0464/

observational biomedical data, yet this uncertainty is rarely taken into account when analyzing the data (e.g. ICD9 codes, from the International Classification of Diseases, Ninth Revision, are often used as proxies for diagnoses), impeding the development of patient-centered technologies.

We aim to demonstrate the utility of data collected in free-living environments from consumer-grade sensors for finding causal relationships. We propose that these data may be able to reproduce findings from highly controlled studies (at lower cost), while also enabling observation of a wider variety of scenarios than observed in the lab (e.g. activity from running to catch a bus or athletic competition). While we aim specifically to infer causes to enable accurate prediction of glucose trends and effective interventions to correct unhealthy highs and lows, there is not yet a method that can weight different observations of a variable differently during inference (based on the certainty of measurements). We hypothesize, though, that this is critical for robust inferences from such data. That is, if we are sure that a particular measurement of glucose indicates hypoglycemia, that measurement should be weighted more strongly in finding causes of hypoglycemia than a measurement that has an equal chance of being within the normal range. Since errors and the values of glucose measurements are not evenly distributed throughout the range, and devices have significant error rates, ignoring this underlying uncertainty of the data can lead to false positives and false negatives. In many cases information on error rates is available, but is not routinely incorporated into inference. For example, studies evaluate accuracy of continuous glucose monitors in various scenarios (e.g. [5,6]) and models of this error have been developed (e.g. [7,8]), leading to the ability to determine which data points may be untrustworthy but not yet the ability to use that information in causal inference.

To handle this, we extend a causal inference method to handle uncertain data, by separating the observation of an event or measurement of a variable's value from the underlying truth of what occurred. This ensures that conclusions drawn from highly uncertain measurements are given less weight than those drawn from reliable measurements and a completely uninformative variable will be treated as missing, instead of propagating errors. The methods are applied to data collected from 17 individuals with T1DM in free-living conditions, using a variety of body-worn sensors. We demonstrate that these patient-generated data can be successfully used to uncover some causes of unhealthy glucose excursions in people with diabetes – but only when uncertainty of the data is properly accounted for.

This paper makes two main contributions: (1) we extend causal inference methods to better handle uncertainty in observational data; (2) we present a novel set of free-living data from people with diabetes that is available for research use (https://idash.ucsd.edu). We demonstrate the utility of the method through rigorous comparison on simulated data and its successful application to the free-living data.

2. Background

2.1. Diabetes

Chronic diseases are rising in prevalence, necessitating tools that can provide feedback directly to individuals, in contrast to the clinician-centered decision-support paradigm. Diabetes in particular affects over 29 million people in the U.S. (over 9% of the population) [9] and the CDC estimates that if current trends continue, this will grow to 33% of the U.S. population by 2050 [10]. The annual cost of diabetes in the U.S., including medical costs and loss of work, was estimated at over $245 billion in 2012, a 41% increase over the 2007 estimate of $174 billion [11]. People with diabetes incur substantial out-of-pocket costs, impeding access to preventative medical services [12]. Diabetes puts patients at risk for many other diseases, and is the primary cause of chronic kidney disease [13], a leading cause of amputations and blindness [14], and a risk factor for heart disease and stroke [15].

T1DM is a chronic autoimmune condition characterized by an inability to produce insulin and is currently an unpreventable, incurable, disease requiring life-long insulin therapy. Since complications are primarily from long-term high or low blood sugar (hyper- or hypoglycemia), they may be preventable with better glucose management [16]. However this requires patients to frequently test their blood glucose and determine whether to administer insulin (to correct high glucose levels) or ingest carbohydrates (to counter falling glucose levels). While glucose has traditionally been measured before and after meals with fingerstick monitors, these discrete samples cannot provide feedback on trends. CGMs instead provide constant feedback on glucose levels, potentially enabling better management [17].

Artificial pancreas systems aim to create closed-loop systems, linking the CGM and insulin pump with control algorithms that regulate glucose without a patient's intervention. These systems have demonstrated increases in glucose within normal ranges (euglycemia) and reductions in hypoglycemia [18], but most tests are conducted overnight in hospitals when glucose control is simpler and they have not been routinely used in ambulatory circumstances where meals, physical activity, and daily life can confound results. Outpatient studies have primarily focused on demonstrating the safety of these systems [19]. The problem is further complicated by the input to the control algorithms: CGMs measure glucose in the interstitial fluid between cells, rather than in blood. This can lead to a delayed signal that is a shifted and transformed version of blood glucose [20], requiring strategies to accurately reconstruct it [21]. Further, while sensors do not have perfect accuracy (due to factors such as noise or calibration errors), understanding real-world performance and factors affecting it has been a significant area of work [5,6]. However, it remains to incorporate what is known about error rates into inference.

Data from body-worn sensors (e.g. heart rate and activity monitors) and mobile apps may provide a more complete picture of the causes of changes in glucose and fill in information gaps. Physical activity can improve glycemic control [22] and increase insulin sensitivity [23], but creates challenges in glucose management since its effects are a function of the activity context (training vs. competition), duration and intensity, and even time of day [24], and physical activity includes both formal exercise and things like running to catch a bus [25]. Recent articles have assessed activity monitors and phone apps in laboratory and free-living conditions, finding that step counts are accurate, but other activity measures may differ from research-grade devices [26,27]. However, it is unclear whether devices are more accurate at tracking changes from a given individual's baseline. More fundamentally, little work has been done to assess causes of hypo- and hyperglycemia in an integrated way in realistic settings (rather than understanding one piece of the puzzle at a time in controlled experiments). To this end, we show that consumer-grade sensors can be used to monitor factors that affect glucose, and with proper analysis can do so accurately enough to find causes of changes in glycemia and eventually, to better inform patients. Further, our publicly available dataset, collected in real-world environments, may enable researchers conducting lab-based studies to understand how these relate to uncontrolled environments and may facilitate computational researchers developing new algorithms to handle the challenges of these data.