Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Samuel Ellis Weinstein April 10, 2023

Vocal Clues to Diabetes Mellitus: Exploring the Ethics and Tech of AI in Clinical Practice

by

Samuel Ellis Weinstein

Melvin Konner
Co-Adviser

Kristin Phillips
Co-Adviser

Anthropology

Melvin Konner
Co-Adviser

Kristin Phillips
Co-Adviser

Vin Tangpricha
Committee Member

John Banja
Committee Member

Pichatorn Suppakitjanusant
Committee Member

2023

Vocal Clues to Diabetes Mellitus: Exploring the Ethics and Tech of AI in Clinical Practice

By

Samuel Ellis Weinstein

Melvin Konner
Co-Adviser

Kristin Phillips
Co-Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Anthropology

2023

Abstract

Vocal Clues to Diabetes Mellitus: Exploring the Ethics and Tech of AI in Clinical Practice
By Samuel Ellis Weinstein

Diabetes mellitus (DM) is one of the most common diseases globally. It incurs enormous economic burdens at the personal, national, and global levels. Unfortunately, inequalities in healthcare systems between health plans and geographical regions and the high direct and indirect healthcare expenses directly impact healthcare access, especially for those in lower economic classes. These factors make access to DM screening challenging and inequitable, contributing to almost one-quarter of all people with diabetes mellitus remaining undiagnosed. Although widely accepted, current standard tests for DM are available only in specially equipped medical centers, are often painful, time-consuming, and can be cumbersome to schedule. Taken together, there is a need for non-invasive methods for diabetes mellitus screening that can be easily accessed. As a sub-study of The Voice Study, which aims to develop a non-invasive and highly accessible screening tool for diagnosing DM using the human voice, my thesis investigates whether the current data of The Voice Study would generate a biased AI algorithm by determining if a variation in voice acoustics exists between sexes and among racial subgroups. The consequences of a biased AI system, specifically in healthcare, would be detrimental as it would result in incorrect diagnoses for specific patient populations, perpetuating and potentially exacerbating already existing disparities in healthcare. Participants included were screened and provided informed consent. Voices were recorded in Emory Healthcare clinics and analyzed using a Computerized Speech Lab with Multi-Dimensional Voice Program software. The 34 parameters of each voice record were analyzed within the subgroups the participant identified. The results display significant variations in multiple parameters between male and female voices. Acoustic variation was also found among racial groups in both male and female subgroups. These results indicate that attention should be given to the diversity of data when designing and deploying AI algorithms, especially in clinical practice. This study should be advanced with more data and analyzed together with other studies to develop specific recommendations for tackling biases in diagnostic AI algorithms, as well as others deployed in clinical practice.

Vocal Clues to Diabetes Mellitus: Exploring the Ethics and Tech of AI in Clinical Practice

By

Samuel Ellis Weinstein

Melvin Konner
Co-Adviser

Kristin Phillips
Co-Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Anthropology

2023

Acknowledgements

Thank you to the Anthropology Department for all their guidance and support throughout my time at Emory and for helping me grow as a student, researcher, and person.

Dr. Paul – who wants nothing but to see his students succeed.
Heather – who works tirelessly to support everyone. You carry this place on your back.
Dr. Konner – who stood by and supported me despite time away.
Dr. Phillips – who took me on last-minute and continuously looked out for me.

Thank you to Dr. Banja for fostering intellectual curiosity and conversations that inspired this research and to the Center for Ethics for being my home at Emory.

Thank you to Dr. Tangpricha for never denying a request for support, despite a (very) busy schedule, and Bright for being the best mentor I could ask for; I am lucky to call you my friend.

Thank you to my family and friends for helping me find the right words.

I am eternally grateful for each of you.

Table of Contents

PREFACE

This thesis explores variations in voice between sexes and among racial subgroups. It is important to acknowledge that the words "race" and "racial" have become increasingly controversial in recent years. While these terms have traditionally been used to describe biological and genetic differences between human populations, it is important to recognize that they are social constructs and not scientifically grounded in biological or genetic fact (Genome.Gov, 2022). Moreover, the use of these terms can perpetuate harmful stereotypes and discrimination (Borrell et al., 2021).

It is also important to clarify that the terms "male" and "female" in this thesis refer to individuals who were born as those sexes, based on biological sex. This distinction is made to differentiate between sex and gender, which are often conflated. Though used interchangeably in some sections of this paper, as quoted by the reviewed literature, gender is a social construct that encompasses a range of identities and expressions, while sex is a biological distinction between males and females (WHO.Int, 2023).

With this in mind, my thesis aims to explore the acoustic variations in voice between sexes and among racial subgroups, while recognizing the complexities and nuances of these social constructs. By examining these variations, we can better understand how they may contribute to biases within a novel AI algorithm that aims to predict diabetes mellitus with voice.

INTRODUCTION

Diabetes mellitus (DM), an umbrella term that refers to both Type 1 Diabetes (T1DM) and Type 2 Diabetes (T2DM), is chronic and one of the most common diseases globally (CDC, 2022). In the most recent release of their atlas of statistics (10th Edition), The International Diabetes Federation (IDF) reported that 537 million adults were living with DM worldwide in 2021, causing 6.7 million deaths that year. The total number of people living with DM is projected to rise to 643 million by 2030 and 783 million by 2045. Diabetes mellitus incurs enormous economic burdens at the personal, national, and global levels. In 2021, at least 966 billion dollars were spent on DM-related health expenditures (IDF, 2023).

In this paper, the term 'diabetes mellitus' is used as an all-encompassing label to refer to both subtypes of the disease; however, Type 1 diabetes and Type 2 diabetes are distinct conditions with different causes, risk factors, symptoms, and treatment options. T1DM is an autoimmune disease that occurs when the immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas, leaving them unable to produce the glucose-regulating hormone. T2DM is a metabolic disorder that occurs when cells become resistant to the effects of insulin and, therefore, cannot use it to regulate blood glucose levels. Beyond clinical predisposition for both conditions, the growing number and high prevalence of DM are due to social, economic, and political factors. Unhealthy diets and high obesity rates, for example, are widely known risk factors for developing T2DM. Further, inequalities in healthcare systems between health plans and geographical regions and the high direct and indirect healthcare expenses directly impact healthcare access, especially for those in lower economic classes (Chawla et al., 2016). These factors make access to screening challenging and inequitable for patients with T1DM and T2DM.

To make matters worse, undiagnosed patients with diabetes mellitus usually remain asymptomatic for years, leaving time for the disease to silently progress with no indication for treatment. A delay in diagnosis and inadequate control of DM is associated with both microvascular and macrovascular diabetic complications (Chawla et al., 2016). Early detection of diabetes mellitus, together with proper control, is essential to alleviate morbidity and mortality from the disease. Unfortunately, due to the aforementioned barriers to screening, almost one-quarter of all people with DM remain undiagnosed (*Statistics About Diabetes | ADA*, 2023).

Currently, the standard screening tests for DM include blood tests for fasting plasma glucose (FPG), hemoglobin A1c (HbA1c, reflecting glucose levels over the past three months), and the 75-g oral glucose tolerance test (OGTT) (CDC, 2023). Although widely accepted as the standard of care, these methods are not without disadvantages. First, they are available only in specially equipped medical centers and must be performed by skilled personnel (Smith, 2022). Only people who can readily access healthcare systems have an opportunity to be screened, while people with limited access do not. Requiring an overnight fast and a two-hour in-person visit, the OGTT can be cumbersome to schedule. Moreover, venipuncture is painful and unpleasant, which can reduce the willingness of individuals at risk to get screened (Smith, 2022). Taken together, there is a need for non-invasive methods for diabetes mellitus screening that can be easily accessed. There have been attempts to develop non-invasive methods for DM screening using techniques such as sparse representation classifiers to analyze human facial block color features, hair elements analysis, dermal electrical resistance, and transdermal sensors or laser irradiation to estimate glucose concentrations in the blood (Heller & Feldman, 2008; Huang et al., 2005; Skladnev et al., 2010). All of these methods, however, have specific limitations and

require further development to be implemented as practical screening tools (B. Zhang et al., 2014).

At the Tangpricha Lab, I have been working alongside my mentor, Dr. Suppakitjanusant, to develop a non-invasive and highly accessible screening tool for diagnosing diabetes mellitus using the human voice since 2021. In the primary study of which this thesis is a substudy of, The Voice Study, voice samples are collected using a portable microphone from healthy individuals and patients with DM and are analyzed alongside the widely accepted diagnostic blood tests to teach an artificially intelligent (AI) system to predict the disease. At the height of the Covid-19 Pandemic, when in-person clinics were transitioned to telemedicine clinics, I began collecting voice samples via a program sent to individuals' cell phones. My abstract published in World Health Organization's COVID-19 Research Database identified the same significant variation in voice acoustics with a cell phone microphone as the professional voice recorder used in The Voice Study, indicating that a cell phone voice recorder may be a feasible tool for DM detection (Weinstein et al., 2022). The possibilities of such a system are tremendous as it can be used to program an application that can easily be made accessible worldwide, quickly predicting DM in the most remote of settings as all that is required is a microphone and the internet. A screening tool as accessible as a cell phone would generate infinite opportunity for diagnosing not just DM for the roughly 135 million people who remain undiagnosed, but the millions of others who suffer from diseases that have the potential to be diagnosed by voice like Parkinson's Disease, Alzheimer's Disease, and Multiple Sclerosis (Pinyopodjanard et al., 2021).

Regardless of the method used to collect the voice, predicting DM or any other disease via this method requires an AI algorithm programmed with vast amounts of voice data paired individually with diagnostic indicators. No matter the quantity of data, though, AI algorithms are

highly prone to biases if the data they are trained on disproportionately represents a specific subgroup and the chosen subgroup significantly differs from other groups in voice acoustics. The consequences of a biased AI algorithm, specifically in healthcare, would be detrimental as it would result in incorrect diagnoses for specific patient populations, perpetuating and potentially exacerbating already existing disparities in healthcare. Thus, it is crucial that special attention be given to the diversity of the data used to develop these systems. In any AI algorithm deployed in healthcare, this is a matter of first identifying any variation among sex, racial, and ethnic subgroups and second, deciding whether the development of a single, universal algorithm is acceptable or if multiple, demographically unique algorithms are necessary to predicate disease without bias.

My thesis aims to investigate whether the current data of The Voice Study would generate a biased AI algorithm. I will regroup the current voice data my mentor, Dr. Pichatorn Suppakitjinasunt, and I collected in Emory Healthcare's outpatient clinics to determine if a variation in voice acoustics exists between sexes and among racial subgroups. By investigating the biological, historical, and anthropological factors that may be responsible for such variation, I will answer the question of whether demographically unique algorithms are necessary to predict DM accurately in all patient populations. Finally, I will provide an ethical framework for implementing AI algorithms into clinical practice.

LITERATURE REVIEW

A. Physiological influence of diabetes mellitus on voice

In its most familiar sense, voice refers to the sounds we produce to orally communicate linguistic information such as meaning, ideas, opinions, and personality. Our ability to adjust the quality of voice, such as by fluctuations in tone and pitch, additionally empowers humans to transmit paralinguistic and nonlinguistic information, such as emotion and age, respectively. Physiologically, voice refers to sound production, or phonation, by vocal fold vibrations that transpire from modulated airflow through the glottis, a valve-like opening between the vocal cords, during breathing. Produced sound propagates through the vocal tract, a region extending between the edge of the lips and nose down to the vocal cords deep in the throat, which contains the oral cavity, nasal cavity, pharynx, and larynx. These voice-modifying structures selectively amplify and attenuate sound at different frequencies via control of the muscular organs contained within (Z. Zhang, 2016). Put simply, phonation is a complex phenomenon involving the generation of airflow and the creation of pressures by the movement of anatomical structures. Its science includes the studies of aerodynamics, acoustics, kinematics and dynamics, and psychoacoustics, and its process requires exquisite coordination between multiple organs and organ systems (Behrman, 2007).

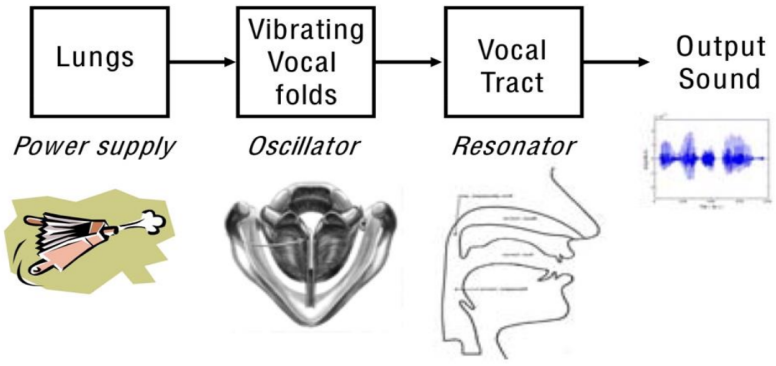

Figure 1. Source-filter theory of speech production (Fant, 1960)

Given the physiology of the vocal tract, it stands to reason that vocal characteristics are directly related to the anatomy of these voice-generating and voice-modifying organs. Pitch, for example, is almost entirely dependent on the length of the vocal cords. Comparatively longer vocal cords in males are precisely why, on average, they have a lower vocal pitch than females. Timbre depends not only on the length and thickness of the vocal folds but also on the specific structure of vocal resonators– the organs involved in amplifying and altering the sounds produced by the vocal cords. The upper resonators are the larynx, pharynx, mouth, nasal, and sinus cavities, and the lower resonators are the trachea, bronchi, and lungs. A deviation from normal of any of these structures, including anomalies caused by systemic diseases, likely affects voice production and quality. Inflammation of the vocal cords due to laryngitis, for instance, makes the voice sound hoarse. Impaired breathing might result in vocal fatigue and aphonia, whereas androgen deficiency can impair the growth of the laryngeal framework during puberty (Abitbol et al., 1999; Gugatschka et al., 2010). Likewise, individuals with neurological conditions like Parkinson's disease or Multiple Sclerosis often report experiencing pitch breaks in their voice, while those with laryngopharyngeal reflux disease may experience dysphonia and a wide range of other vocal-related issues (Hamdan et al., 2012).

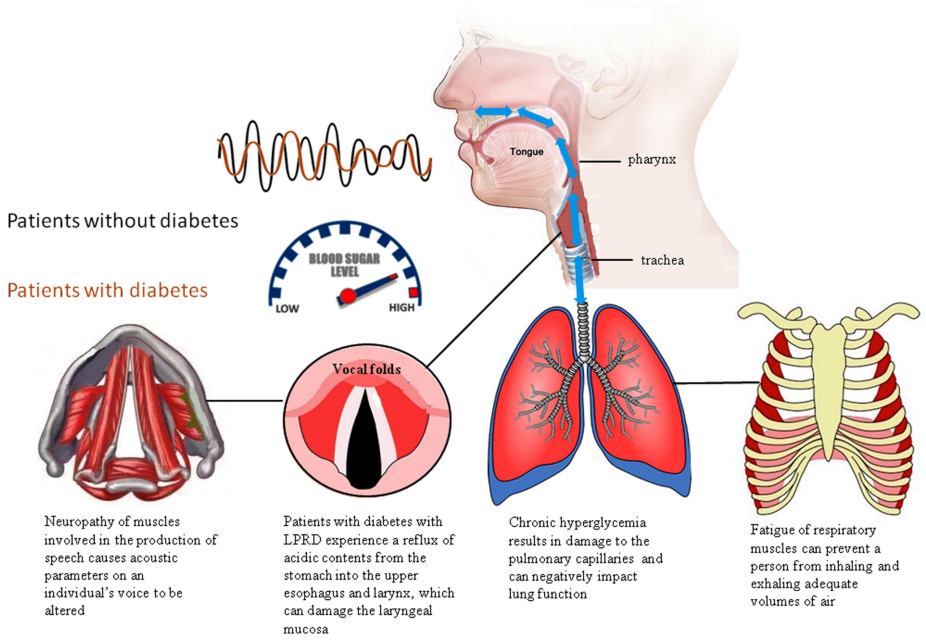

Figure 2. DM physiological changes that affect speech production (Saghiri et al., 2022)

Very little research exists that examines the relationship between voice quality and DM. A scoping review of literature published between March 2000 and September 2021 found only nine studies relevant to the effect DM has on voice quality or the applications of such a correlation to future technology (Saghiri et al., 2022). Of the nine articles examined, only five contained clinical studies, and none of them analyzed patients with Type 1 Diabetes, signifying the need for additional research on this subject. The comprehensive review of literature presented in this paper will examine each of the articles reviewed by Saghiri et al., those reviewed by Ravi & Gunjawate in the only other systemic review of literature related to the effect of DM on voice (Ravi & Gunjawate, 2019), as well as additional relevant articles published before March 2023.

DM is a chronic disease that affects how the body metabolizes glucose, the primary source of energy for humans (IDF, 2023). Type 1 Diabetes (T1DM) occurs when the pancreas cannot produce insulin, and Type 2 Diabetes (T2DM) occurs when cells become resistant to

insulin and, therefore, cannot use it. In either case, DM prevents cells from consuming and using glucose, leading to increased glucose concentration in the blood (hyperglycemia). Both the cells diminished capability of metabolizing glucose and the hyperglycemia that results are the two primary sources of all disease complications. While treatments currently available for DM, such as supplemental insulin and medications that improve the action of insulin, are highly effective mediators of the development of disease complications, DM affects almost every major body system, even in individuals with the most well-controlled diabetes, as glucose is required to fuel respiration in virtually every cell.

Chronic hyperglycemia as a result of DM can damage the walls of the blood vessels that supply cells with oxygen and nutrients. In the cardiovascular system, DM is commonly associated with comorbidities such as high blood pressure (hypertension), high cholesterol, and coronary artery disease, all of which may lead to fatal complications such as heart attack and stroke. Damage to small blood vessels in the kidneys may make them less efficient or cause them to fail altogether. Further, diabetic neuropathy, or nerve damage as a result of DM, may not only cause numbness, tingling, and pain in the hands, feet, and legs, but can also affect the autonomic nerves that control the internal organs, leading to problems with digestion, sexual function, and hormone regulation. Regardless of whether or not complications develop or how severe they may become, DM has a profound impact on the physiology of all cells, organs, and body systems.

Phonation by the phonatory apparatus, a collection of all organs involved in sound production from the tongue to the lungs, depends entirely on the interplay of numerous muscles, nerves, and vasculature, all of which are highly sensitive to and affected by the pathophysiologies of DM, as well as its associated comorbidities. Thus, DM likely affects the anatomical structure and physiology of these organs. A change in blood glucose level might

affect the elastic properties of involved tissues, which consequently may result in changes in spectral characteristics of the human voice, significantly impacting voice quality (Hamdan et al., 2012). Diabetic neuropathy can cause damage to the nerves that control the muscles used in voice production, leading to weakness, hoarseness, or other changes in voice quality. Dehydration due to DM can dry out the vocal cords, making them less flexible, and acid reflux, a common problem for people with DM, can lead to irritation and inflammation of the vocal cords. Other comorbidities caused by DM, such as hormone imbalances and diabetic myopathy, may additionally impact voice quality.

One of the earliest attempts to identify a difference in vocal quality between patients with Type 2 Diabetes and healthy individuals was conducted in a study performed by Hamdan et al. In this patient-reported outcome study, 105 patients with Type 2 Diabetes and 33 healthy individuals were surveyed about their phonatory symptoms. Hamdan et al. found a significantly higher presence of vocal strain and hoarseness in subjects with diabetes than in those without (Hamdan et al., 2013). Although this study offers a certain degree of understanding about the link between DM and voice, its findings rely entirely on subjective reports of phonatory symptoms; it lacks the essential empirical evidence needed to arrive at more definitive conclusions.

In another study by Hamdan et al., a more direct approach was taken toward measuring vocal parameters. Researchers utilized the GRABS (grading, roughness, asthenia, breathiness, and straining) scale to evaluate different aspects of voice quality. The study found that 9% of the subjects with T2DM had moderate-grade roughness, while no one from the control group exhibited this symptom. The authors also examined subgroups of subjects with diabetes and found that those with poor glycemic control and neuropathy had higher scores for vocal grading

and straining (Hamdan et al., 2012). These initial findings suggest that there could be a connection between diabetes, neuropathy, and changes in voice quality.

A preliminary analysis of The Voice Study published in the *Journal of Voice* in 2021 (Pinyopodjanard et al., 2021) examined 83 patients with DM and 70 healthy controls. Voice parameters, including fundamental frequency (Fo), jitter, shimmer, amplitude perturbation quotient, noise-to-harmonic ratio, smoothed amplitude perturbation quotient, and relative average perturbation, were analyzed using a Computerized Speech Lab with the Multi-Dimensional Voice Program. The study found that Voice fundamental frequency was significantly associated with diabetes when controlled for age, BMI, presence of hypertension, and dyslipidemia, particularly in females.

B. Acoustic variation among demographic populations

The human voice serves as a primary source of communication between individuals; it is one of the most natural, energy-efficient ways of interacting with each other. The voice contains various information and plays a fundamental role in social interaction by allowing us to share insights about our emotions, fears, feelings, and excitation by modulating its tone or pitch (Fagherazzi et al., 2021).

Phonation involves the interplay and activation of numerous biomechanical, neurological, and physiological systems. In addition to these systems' high sensitivity to disease, causing detectable variations, voice is entirely unique to each individual (Podesva & Callier, 2015), containing information about the speaker's development, gender, age, language, culture, and emotional state (Andrianopoulos et al., 2001). Further, racial and gender differences have been documented in voice characteristics, which raises concerns about whether significant differences

exist between subgroups, especially when voice is employed in artificial intelligence (AI). Because the outputs of AI models are derived from the data used to train them, the input data, an algorithm trained on a disproportionally excessive amount of one kind of data would output results that disproportionately advantage the excessive input. This leaves AI algorithms highly prone to bias (AI bias). In healthcare, exceptional attention must be given to the training data, as even minuscule biases can be detrimental. A biased voice and AI algorithm for diagnosing DM, for example, would result in incorrect diagnoses for specific patient populations, leading to the mistreatment of or even no treatment of the disease. This is not to mention that biased AI algorithms in healthcare would exacerbate already existing disparities. Combatting AI bias in The Voice Study means first identifying if a variation exists between sexes and among subgroups and, second, examining how such a difference would impact the AI algorithm when used to predict disease. Based on the results, it can be determined whether developing a single, universal algorithm is acceptable or if multiple, demographically unique algorithms are necessary to predicate disease unbiasedly.

Given that racial and gender differences have been documented in inherent voice characteristics (Chen et al., 2022), these differences can result in crucial bias issues when voice is used to train an AI algorithm, whether it be used for biometric security or to predict disease, especially when deployed on a large scale. Therefore, it is critical to examine if sex and racial subgroups have inherent differences in voice characteristics, then causing biases in AI algorithms. It is hypothesized that if any variation exists among racial subgroups, it would be due to slight differences in genetic makeup, giving rise to anatomical variations of the phonatory apparatus. Few studies exist, however, that identify such anatomical uniqueness among racial

subgroups. While the biological origins of acoustic variation remain relatively unexplored, multiple studies identify quantifiable variations among racial subgroups.

A study published in the *Journal of Clinical Linguistics and Phonetics* (Awan & Mueller, 1992) studied speech samples from groups of Caucasian, African American, and Hispanic kindergarten-age children and compared measures of mean speaking fundamental frequency (Fo), maximum and minimum speaking fundamental frequency (Fhi and Flo, respectively), standard deviation, and speaking range. The results of these comparisons indicate that Caucasian, African American, and Hispanic kindergarten-age children do not differ significantly in terms of mean speaking Fo.

In another study published in *The Journal of the Acoustical Society of America,* Deutsch et al. suggests that pitch differences may be due to cultural rather than racial factors (Deutsch et al., 2009). Different linguistic communities have expectations for their members, and an individual's vocal characteristics can adapt to the surrounding community, even sounding similar in pitch. Deutsch et al. also examined the pitch levels of females from two Chinese villages, each community being homogenous both ethnically and culturally, with the dialects of Mandarin spoken in the villages also being similar. The Fo values were clustered within each village but differed by approximately three semitones. These data support the claim that Fo is influenced by a representation acquired through long-term exposure to the speech of others and suggests a cultural, rather than a physiological, influence on pitch.

In contrast, a physiological study by Xue and Hao attempted to determine vocal tract measurements of speakers from different races (Xue & Hao, 2006). The study included 120 Caucasians, African Americans, and Chinese subjects. Males and females were separated within each racial subgroup. Race was found to be a significant variable for oral volume and total vocal

tract volume. The Caucasian American speakers had significantly smaller mean volume and mean total tract volume than Chinese speakers. There were no significant differences in the oral volume and total tract volumes between Caucasian American and African American speakers. Within the male speakers, Chinese males had significantly larger oral volume than both Caucasian American and African American male speakers. Further, Chinese male speakers had significantly larger total vocal tract volumes than Caucasian and African American male speakers. Within the female group, results displayed that Chinese female speakers had significantly shorter vocal tract lengths than the other two groups. Caucasian female speakers had significantly larger pharyngeal volume than African American and Chinese female speakers. The authors concluded that such a variation in pitch is likely due to physiological differences between the racial subgroups.

Differences in speaking fundamental frequency, vocal quality, volume, and noise-related measurements have been reported among cultures, races, and native languages; however, data are sparse and lack consistency concerning test and diagnostic protocols and test standardization. The data collected in The Voice Study, however, is vast, demographically diverse, and collected and analyzed following a strict, consistent protocol. The present study offers a protocol for the assessment and analysis of acoustic variation between sexes and among racial subgroups that is identical to the protocol of the study it supports, The Voice Study, removing concerns for consistency in voice collection, assessment, and analysis. Thus, an analysis of acoustic variation among demographic populations within The Voice Study would be the best way to determine if demographically unique algorithms are necessary to predict DM unbiasedly.

C. Ethical considerations of artificial intelligence in healthcare

Artificial intelligence (AI) is a field that combines computer science with robust data sets to problem-solve and perform tasks that would normally require human intelligence (*Artificial Intelligence | IBM*, 2023). It involves the development of algorithms and computer programs that can analyze data, learn from that data, and make decisions or take actions based on the insights derived from it. Described as the "fourth industrial revolution" by the World Economic Forum (*The Fourth Industrial Revolution*, 2016), AI is already changing the way we learn, perform tasks, practice medicine, and even drive cars as it becomes increasingly prominent in our daily lives (Ouchchy et al., 2020). The Forum claimed: "We stand on the brink of a technological revolution that will fundamentally alter how we live, work, and relate to one another. In its scale, scope, and complexity, the transformation will be unlike anything humankind has experienced before" (*The Fourth Industrial Revolution*, 2016). Given the revolutionary nature of AI and the fact that the technology is currently being developed and deployed in virtually all sectors, ethics has been identified as a priority concern (Gibney, 2020; Ouchchy et al., 2020). When discussing the implementation of AI into a field as trust-based and deeply personal as healthcare, the vast and complex issues related to AI make ethics an indispensable consideration (Murphy et al., 2021).

Within the healthcare field, the promise of AI lies in its proven ability to promote healthy behaviors, detect and intervene early in infectious illnesses and environmental health threats, and prevent, diagnose, and treat disease (Bhagyashree et al., 2018; Johnson & Pauwels, n.d.; X. Zhang et al., 2017). While it is quite evident that AI holds the potential to improve health and health systems, an analysis of 103 articles related to the ethics of the matter in a scoping review of literature (Murphy et al., 2021) suggests that its introduction should be approached with

caution. In their study, Murphy et al. identified four common ethical themes across the health applications of AI addressed in the literature: data privacy and security, trust in AI, accountability and responsibility, and bias. As it applies to The Voice Study, which aims to develop an AI algorithm to predict diabetes mellitus equitably, this section focuses specifically on the implications of AI bias that would arise should the predictive DM model be trained on disproportionate data.

AI bias refers to the systematic error or unfairness in decision-making algorithms that result from the unintentional adoption or reinforcement of prejudices or stereotypes. These biases can occur in various ways throughout the development and deployment of artificial intelligence algorithms, including the data used to train the model, the selection of features used in the model, and the algorithm's decision-making process (Parikh et al., 2019). To prevent the infiltration of bias into AI algorithms, each potential source must be carefully considered. Focused on investigating whether the current data of The Voice Study would generate a biased AI algorithm, my thesis examines the data currently being used to train the model for acoustic variation between racial subgroups. If a difference in voice exists, special attention must be given to the representation of each subgroup within the dataset, its balance, and whether separate algorithms are required to predict disease in each subgroup.

The prevailing concern with the development of AI algorithms is that they are developed by humans who, by nature, are fallible and subject to their own values and implicit biases (Murphy et al., 2021). As described by the literature reviewed by Murphy et al., these values often reflected those that are societally endemic and, if carried into the design of AI algorithms, could consequently produce outputs that advantage certain population groups over others (Murphy et al., 2021). Similarly, bias manifested in datasets if the data used to train AI

algorithms were inaccurate, incomplete, or unrepresentative of the population they aimed to target, rendering them ungeneralizable (Murphy et al., 2021). Such bias in datasets has been noted by Murphy et al. to potentially perpetuate systemic inequities based on race, gender identity, and other demographic characteristics, which may limit the performance of AI as a diagnostic and treatment tool (Murphy et al., 2021). Suggestions to address AI bias, offered by *The AI for Good Global Summit* in Geneva, Switzerland, in 2017, included building AI algorithms that reflect current ethical healthcare standards and ensuring a multidisciplinary approach to AI design and deployment (Murphy et al., 2021).

METHODS

All voices analyzed in the present study were collected as part of The Voice Study and therefore follow an identical protocol. Methods section *A. Voice Study Recruit* describes the entire protocol for voice collection and acoustic analysis of The Voice Study. Methods section *B. Demographic Regrouping* describes the secondary protocol for analysis of acoustic variation between sexes and among demographic populations of the voice data collected in The Voice Study. Considering such a discrepancy in the field, and given the tremendous potential of The Voice Study, it is essential that acoustic variation be measured between sexes and among racial subgroups within the study, using identical acoustic assessment and analytic methods to predict disease in an unbiased and equitable manner.

A. Voice Study Recruit

i. Participants

The Voice Study is an ongoing, prospective study of 300 individuals recruited at Emory Healthcare's clinics as of March 2021. The study is approved by the Emory University Institutional Review Board and is conducted under the guidance of Dr. Vin Tangpricha. The 300 participants included are divided into three groups according to the advancement of disease. Patients with DM with an HbA1c level of more than 7% are classified into Poor-Controlled DM (Group 1). Patients with an HbA1c level equal to or less than 7% are classified into Well-Controlled DM (Group 2). Individuals without DM are classified as Healthy (Group 3). Voice data will only be collected to fill each of the three groups equally (100 subjects in Poor-Controlled DM, 100 subjects in Well-Controlled DM, and 100 subjects in the Healthy Group).

ii. Protocol

Participants who met the criteria for screening were interviewed using a predefined questionnaire to collect demographic information and the duration of disease. Relevant medical information, including results from the OGTT, hemoglobin A1c test, and serum glucose test, were extracted from the electronic medical record and the participant questionnaire. The glycemic control was described based on the HbA1c level. All participants provided signed informed consent before participating in the study. The inclusion criteria for both DM Groups, Poor-Controlled DM and Well-Controlled DM, are shown in Table 1. The Healthy Group (Group 3) criteria included all inclusion and exclusion criteria except for DM diagnosis.

Inclusion Criteria	**Exclusion Criteria**		
Patient is 18 to 70 years old at time of screening	Patient is pregnant or breastfeeding	Active smoking or smoking cessation within 6 months	Patient has a history of illicit drug abuse
Patient has been diagnosed with Type 2 Diabetes in the past 5 years based on: HbA1c **OR** Fasting Blood Sugar Test (OGTT) **OR** is currently taking DM drugs	Patient had an acute myocardial infarction or stroke within the past 6 months, or has uncontrolled hypertension (systolic blood pressure > 140 mmHg or diastolic blood pressure > 90 mmHg)	Patient has a speech disorder or history of abnormal voice within the previous 2 weeks	Patient has neurological or active mental disorders
Patient is able to sustain the vowel /:a/ at a comfortable pitch and loudness for 5 seconds	Patient has a history of thyroid disease	Patient is currently enrolled in another conflicting or interventional study	In opinion of the study team, the patient is too sick to participate

Table 1. Inclusion and Exclusion Criteria for patients with DM enrollment into The Voice Study

iii. Glucose measurements

Point-of-care blood glucose testing was conducted at the participant's clinic visit using the OneTouch Verio Flex glucometer. Capillary glucose measurement was performed by

pricking a sterile lancet on one of the fingertips to obtain a drop of blood for use in the glucometer.

iv. Voice recording

During an outpatient clinic appointment, the participant was requested to produce the sustained vowel at the usual frequency and strength of /a:/ after taking a deep breath in order to obtain maximum phonation time without using expiratory reserve air. This vowel sound was chosen per The American Speech-Language-Hearing Association's recommendation for standardized clinical auditory-perceptual assessment of voice, outlined in the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). Additionally, the /a:/ sound can be continuously vocalized and has the most occurrence compared to other syllables (Chen et al., 2022), therefore being feasible and adequate for voice biometrics. Further, the voice biometric is practical and has high user acceptance in real applications (Poddar et al., 2018). Recordings were performed while participants were seated in a silent room with a head-mounted mouth microphone positioned 4 centimeters from the lips and 45 to 90 degrees from the front of the mouth. All voice datasets were amplified and analog-to-digital conversed via USBPre2. Using the voice recorder software Garageband, voice samples were saved as *.wav* files. Initial 3.5 seconds of vowel /a:/ emission was collected for acoustic voice analysis; the commencement of emission was omitted to prevent interference from voice attachment on data processing.

v. Voice acoustic assessment

Acoustic parameters were analyzed using Computerized Speech Lab (CSL) model 4500 by Dr. Nittaya Kasemkosin in her voice lab at Mahidol University in Bangkok, Thailand. CSL is considered the gold standard system for acoustic analysis and has been validated to assess voice pathology in many controlled trials (Bielamowicz et al., 1996; Burris et al., 2014; Read et al.,

1992). The Multi-Dimensional Voice Program (MDVP), a computerized voice analysis system, in conjunction with CSL, is a versatile voice-processing and spectrographic analysis software package ideally suited for use in determining the primary acoustic parameters, including fundamental frequency (Fo), fundamental frequency variation (vFo), jitter, shimmer, relative average perturbation (RAP), noise-to-harmonic ratio (NHR), as well as other secondary parameters.

Fo, measured in Hertz, is defined as the number of times a sound wave produced by the vocal cords repeats during a given time period. It is also the number of cycles of opening/closure of the glottis. vFo, measured in %, represents the relative standard deviation of the period-to-period calculated Fo. Measurement of Fo, disturbance, jitter, and shimmer, have proven useful in describing vocal characteristics. Jitter, measured in %, refers to frequency variation from cycle to cycle. It is affected mainly by the lack of control of the vibration of the vocal cords. Shimmer, measured in %, relates to the amplitude variation of the sound wave. It changes with the reduction of glottal resistance and mass lesions on the vocal cords and is correlated with the presence of noise emission and breathiness. RAP, measured in %, is a quantitative measure of the voice on a sustained vowel. It minimizes the effects of slow changes in the fundamental frequency by averaging across three successive vibratory cycles to provide an estimate of the value of the middle period had there been no perturbation present. NHR is a general evaluation of noise present in vocalization. It is an assessment of the ratio between non-periodic components, which are the glottal noise, and periodic components, which are the vibration of the vocal cords. The quality of the recording of each individual's sound, namely the sustained /a:/, was blindly evaluated by a speech therapist.

vi. Statistical analysis

All study data were checked for normality and presented as mean if they were normally distributed or median if they were not normally distributed. For baseline characteristic data, the independent t-test was used to compare continuous variables, and the Chi-square test was used to compare categorical variables between groups. Multivariate and logistic regression analyses were performed to identify independent variables associated with CFRD. A two-tailed p-value of less than 0.05 was considered statistically significant. Statistical analyses were performed using Excel version 16.70.

B. Demographic Regrouping

The analysis of demographic variation included all voices previously collected in The Voice Study, totaling 135. As such, the methods for this substudy were identical, including participant selection, the study protocol, voice recording, and voice acoustic assessment. The following regrouping of data collected in the voice study was performed to determine if a significant variation exists between sexes and among racial subgroups. Data were first sorted to analyze acoustic variation between male and female participants. A Two-tailed T-test was conducted on each of the 34 parameters outputted by the MDVP, with all male participants being sample one and all female voices being sample two. A p-value of less than 0.05 was considered significant.

Considering the results of these tests and that prior research has established that the voices of males and females have significant differences, these groups were kept separate to analyze acoustic variation among racial subgroups. Within each of the two sex groups, data were further divided into three subgroups: African American, Caucasian, and All Others. Due to the

insufficient amount of data within The Voice Study required to form the individual groups, Asian, Indian, and Hispanic voice data of subjects so identifying were combined into the group named ‘All Others.’ A Single Factor Analysis of Variance (ANOVA) Test was performed on each of the 34 parameters, comparing each of the three racial subgroups within each sex, totaling 64 ANOVA tests. A p-value of less than 0.05 was considered significant.

RESULTS

Table 2 defines each of the 34 parameters outputted by the MDVP. The p-values of the Two-tailed T-tests performed on each parameter to compare 62 male, and 73 female voices are displayed in Table 3. Of the 34 parameters, twelve have significant differences between sexes, as evidenced by a p-value of less than 0.05. Of these significant parameters, females had higher mean values for Fo, Mfo, Fhi, Flo, STD, Jitt, RAP, PPQ, DSH, NSH, and PER compared to males and a lower mean value for To.

Significant variations were also found among the three racial subgroups in each of the parameters displayed in Table 4. The Female subgroup had 44 individuals identifying as African American, 25 individuals identifying as Caucasian, and four individuals classified into the All Others subgroup. The Male subgroup had 27 individuals identifying as African American, 28 individuals identifying as Caucasian, and seven individuals classified into the All Others subgroup. The ANOVA Test does not inform which groups vary from each other, only if at least one difference exists. ANOVA tests for each significant parameter are displayed in Tables 5-9.

Abbreviation	Description
Fo (Hz)	Average fundamental frequency for the vocalization.
Mfo	Mean fundamental frequency.
To	Average Pitch Period is the average value of all extracted using the VOICE command pitch period values. Voice break areas are excluded.
Fhi (Hz)	Highest fundamental frequency in the vocalization.
Flo (%)	Lowest fundamental frequency in the vocalization.
STD	Standard deviation of the fundamental frequency in the vocalization.
PFR	Phonatory fundamental frequency range in semitones.
Fftr (Hz)	The frequency of the most intensive low-frequency Fo-modulating component.
Fatr (Hz)	The frequency of the most intensive low-frequency amplitude-modulating component.
Tsam	Length of analyzed voice data sample.
Jitta (µs)	Absolute jitter gives an evaluation in microseconds of the period-to-period variability of the pitch within the analyzed voice sample.
Jitt (%)	Gives an evaluation of the variability of the pitch period within the analyzed voice sample in percent. It represents the relative period-to-period (very short-term) variability.
RAP (%)	Relative average perturbation gives an evaluation of the variability of the pitch period within the analyzed voice sample at smoothing factor 3 periods.
PPQ (%)	Pitch perturbation quotient gives an evaluation in percent of the long-term variability of the pitch period within the analyzed voice sample at smoothing factor 3 periods.
sPPQ (%)	Smoothed PPO gives an evaluation in percent of the long-term variability of the pitch period within the analyzed voice sample at smoothing factor 55 periods. sPPQ correlates with the intensity of a frequency tremor.
vFo (%)	Fundamental frequency variation represents the relative standard deviation of the period-to-period calculated FO. It reflects the very long-term variations of FO within the analyzed voice sample.
ShdB (db)	Absolute shimmer gives an evaluation the peak-to-peak amplitude.
Shim (%)	Shimmer percent gives an evaluation in percent of the variability of the peak-to-peak amplitude within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability of the peak-to-peak amplitude.
APQ (%)	Amplitude perturbation quotient gives an evaluation in percent of the variability of the peak-to-peak amplitude within the analyzed voice sample at smoothing factor 11 periods.
sAPQ (%)	Smoothed APO gives an evaluation in percent of the long-term variability of the peak-to-peak amplitude within the analyzed voice sample at smoothing factor 55 periods. sAPQ correlates with the intensity of an amplitude tremor.
vAm (%)	Peak amplitude variation represents the relative standard deviation of the period-to-period calculated peak-to-peak amplitude. It reflects the very long-term amplitude variations within the analyzed voice sample.
NHR	Noise-to-harmonic ratio is an average ratio of energy of the in-harmonic components in the range 1500-4500 Hz to the harmonic components energy in the range 70-4500 Hz. It is a general evaluation of the noise present in the vocalization.
VTI	Voice turbulence index is an average ratio of the spectral in-harmonic high-frequency energy to the spectral harmonic energy in stable phonation areas. VTI measures the relative energy level of high-frequency noise, such as turbulence.
SPI	Soft phonation index is an average ratio of the lower-frequency to the higher-frequency harmonic energy. This index is not a measurement of abnormality but rather a measurement of the spectral "type" of the vocalization.
FTRI (%)	Fo-tremor intensity index shows (in percent) the ratio of the frequency magnitude of the most intensive low-frequency modulating component (Fo-tremor) to the total frequency magnitude of the analyzed voice signal.
ATRI (%)	Amplitude tremor intensity index shows (in percent) the ratio of the amplitude of the most intensive low-frequency amplitude-modulating component (amplitude tremor) to the total amplitude of the analyzed voice signal.
DVB	Degree of Voice Breaks is the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analysed part of the signal
DSH (%)	The Degree of Subharmonics measures the percentage of time the subharmonic is present.
DUV (%)	Degree of Voiceless measures the ability of the voice to sustain uninterrupted voicing.
NVB	Number of times the fundamental period was interrupted during the sample.
NSH	Number of autocorrelation segments where the pitch was found to be a sub-harmonic of Fo.
NUV	Number of Unvoiced Segments detected during the autocorrelation analysis.
SEG	Total number of segments computed during the autocorrelation analysis.
PER	Number of pitch periods detected during the voice sample.

Table 2. 34 acoustic parameters extracted by the MDVP software analysis (Behrman, 2007)

Parameter	Male Mean	STDEV (M)	Mean Female	STDEV (F)	p-Value
Fo	120.85	24.01	188.50	45.28	0.00000
Mfo	120.76	24.01	187.94	45.96	0.00000
To	8.57	1.52	5.64	1.49	0.00000
Fhi	128.90	33.79	210.27	57.05	0.00000
Flo	115.53	24.09	176.47	44.88	0.00000
STD	1.86	2.13	4.01	5.83	0.00681
PFR	2.77	2.80	4.07	4.95	0.07017
Fftr	4.04	1.84	4.78	2.49	0.08833
Fatr	4.44	2.10	3.82	1.80	0.13358
Tsam	2.39	0.67	2.32	0.66	0.58045
Jitta	79.92	79.57	78.97	82.89	0.94616
Jitt	0.92	0.80	1.35	1.22	0.01873
RAP	0.54	0.49	0.80	0.70	0.01364
PPQ	0.54	0.48	0.79	0.75	0.02699
sPPQ	1.03	0.92	1.25	2.11	0.44467
vFo	1.57	1.80	2.38	4.29	0.16635
ShdB	0.57	0.38	0.48	0.31	0.13614
Shim	6.43	4.19	5.30	3.27	0.07982
APQ	4.79	2.73	3.89	2.60	0.05049
sAPQ	7.23	4.20	6.19	3.33	0.11383
vAm	13.28	6.27	14.18	7.39	0.45202
NHR	0.17	0.06	0.17	0.07	0.97601
VTI	0.07	0.04	0.06	0.05	0.32349
SPI	11.38	10.12	11.85	8.41	0.77120
FTRI	0.54	0.50	0.45	0.39	0.26870
ATRI	5.47	3.18	5.66	4.66	0.81327
DVB	0.38	2.07	0.61	2.74	0.59954
DSH	0.37	1.36	2.37	5.13	0.00337
DUV	5.55	12.49	3.73	13.35	0.41614
NVB	0.06	0.31	0.19	0.88	0.27875
NSH	0.31	1.21	1.71	3.63	0.00416
NUV	3.92	8.85	3.00	11.37	0.60587
SEG	79.08	22.25	76.95	21.83	0.57550
PER	284.53	95.68	437.85	199.91	0.00000

Table 3. Mean, Standard Deviation (STDEV), and Two-tailed T-test P-values of all parameters with significant variation between male and female subgroups

Females:	**SEG** (p-value = 0.00016)	**Tsam** (p-value = 0.00015)	**PER** (p-value = 0.00032)
Males:	**NSH** (p-value = 0.00083)	**DSH** (p-value = 0.00051)	

Table 4. ANOVA Test P-values of all parameters with significant variation between demographic subgroups (including individuals with and without DM)

Anova: Single Factor, FEMALE
Parameter: **SEG**
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
African American	44	3079	69.9772727	282.487844
Caucasian	25	2101	84.04	555.54
All Others	4	437	109.25	423.583333

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7569.09355	2	3784.54677	9.90323245	0.00016318	3.1276756
Within Groups	26750.6873	70	382.152675			
Total	34319.7808	72				

Table 5. Analysis of Variance (ANOVA) of **SEG** parameter in the female subgroup

Anova: Single Factor, FEMALE
Parameter: **Tsam**
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
African American	44	92.951	2.11252273	0.25303272
Caucasian	25	63.443	2.53772	0.50046229
All Others	4	13.178	3.2945	0.390015

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.87709531	2	3.43854766	10.0034439	0.00015093	3.1276756
Within Groups	24.061547	70	0.34373639			
Total	30.9386423	72				

Table 6. Analysis of Variance (ANOVA) of **Tsam** parameter in the female subgroup

Anova: Single Factor, FEMALE
Parameter: **PER**
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
African American	44	17224	391.454545	21499.463
Caucasian	25	11604	464.16	35205.9733
All Others	4	3135	783.75	172493.583

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	590604.323	2	295302.162	9.03893572	0.00032218	3.1276756
Within Groups	2286901.02	70	32670.0146			
Total	2877505.34	72				

Table 7. Analysis of Variance (ANOVA) of **PER** parameter in the female subgroup

Anova: Single Factor, MALE
Parameter: **NSH**
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
African American	27	4	0.14814815	0.59259259
Caucasian	28	2	0.07142857	0.06878307
All Others	7	13	1.85714286	8.80952381

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	19.0557262	2	9.52786312	8.01669068	0.00083201	3.15312326
Within Groups	70.1216931	59	1.18850327			
Total	89.1774194	61				

Table 8. Analysis of Variance (ANOVA) of **NSH** parameter in the male subgroup

Anova: Single Factor, MALE
Parameter: **DSH**
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
African American	27	4.878	0.18066667	0.881292
Caucasian	28	2.804	0.10014286	0.13657035
All Others	7	15.12	2.16	10.0368143

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	25.4353054	2	12.7176527	8.64230919	0.00051079	3.15312326
Within Groups	86.8218774	59	1.47155724			
Total	112.257183	61				

Table 9. Analysis of Variance (ANOVA) of the **DSH** parameter in the male subgroup

DISCUSSION

The results identify significant variations between male and female subgroups within the data collected in The Voice Study, as displayed by p-values of less than 0.05 in twelve of the 34 parameters. These results align with previous research that informs of acoustic variation between sexes. In the second regrouping of data that analyzes variance among African American, Caucasian, and All Others subgroups, variations were also found.

Within the Female subgroup, significant variation among subgroups was found in SEG (the total number of segments), Tsam (the length of analyzed voice data sample), and PER (the number of pitch periods) parameters. Within the Male subgroup, significant variations among groups were found in NSH (the number of autocorrelation segments where the pitch was found to be a sub-harmonic of Fo) and DSH (the percentage of time the subharmonic is present) parameters. Though all five of these parameters are considered secondary and not commonly studied in literature (Pinyopodjanard et al., 2021), there is at least some variance among racial subgroups that must be must be considered. When talking about using this data to train an AI algorithm that will be used to predict DM with voice, the variation in acoustics among racial subgroups may contribute to the formation of bias if not accounted for. Thus, special attention must be given to the diversity and balance of data used to train AI algorithms. These algorithms should also be designed to be aware and cautious of the implications such acoustic variations may cause.

The results presented identify significant variations in voice between sexes and among racial subgroups that were not previously found; however, this study was not conducted without limitations. First, there was not enough data collected to form individual subgroups for participants identifying as Hispanic, Indian, and Asian. Further, combining the few voice data of

the included participants may distort the findings, and the All Others subgroup needed more data to claim true statistical significance. Second, the accuracy of the voice data may be affected by several factors, such as the quality of the recording, the microphone used, and the surrounding noise. These factors may lead to errors in the analysis of the voice data.

Most limiting, however, is the need for more consistency and test standardization in the field concerning voice collection methods, acoustic assessment and analysis, and detailed protocols. Because acoustic measurements are so sensitive to the method by which they are retrieved (Pinyopodjanard et al., 2021), voice data collected and analyzed following identical protocols can be compared. About The Voice Study, this study analyzes demographic acoustic variations confidently, without concerns for the inconsistencies mentioned above, and its protocol can be replicated in the future when more data become available to draw more robust conclusions. At that stage of The Voice Study, acoustic variation among demographic populations should also be analyzed directly by the AI algorithm for its potential implications by running it through the system to test whether the outputs predict DM accurately in each demographic population.

This study should be analyzed together with other studies to develop specific recommendations for tackling biases in diagnostic AI algorithms. Additional studies should be conducted to explore acoustic variation among native language, ethnicity, and other demographics, not just within The Voice Study, but for all AI systems that use voice as a diagnostic tool. The same special attention should be given to any potential variation among demographic groups for all AI systems, especially those deployed in clinical practice.

REFERENCES

Abitbol, J., Abitbol, P., & Abitbol, B. (1999). Sex hormones and the female voice. *Journal of Voice: Official Journal of the Voice Foundation*, *13*(3), 424–446. https://doi.org/10.1016/s0892-1997(99)80048-4

Andrianopoulos, M. V., Darrow, K., & Chen, J. (2001). Multimodal Standardization of Voice Among Four Multicultural Populations Formant Structures. *Journal of Voice*, *15*(1), 61–77. https://doi.org/10.1016/S0892-1997(01)00007-8

Artificial Intelligence | IBM. (2023). Ibm.Com. https://www.ibm.com/topics/artificial-intelligence

Awan, S. N., & Mueller, P. B. (1992). Speaking fundamental frequency characteristics of centenarian females. *Clinical Linguistics & Phonetics*, *6*(3), 249–254. https://doi.org/10.3109/02699209208985533

Banja, J. D. (2019). *Patient Safety Ethics: How Vigilance Mindfulness Compliance & Humility Can Make Healthcare Safer*. John's Hopkins University Press.

Behrman, A. (2007). *Speech and Voice Science*. Plural Publishing, Inc.

Bhagyashree, S. I. R., Nagaraj, K., Prince, M., Fall, C. H. D., & Krishna, M. (2018). Diagnosis of Dementia by Machine learning methods in Epidemiological studies: A pilot exploratory study from south India. *Social Psychiatry and Psychiatric Epidemiology*, *53*(1), 77–86. https://doi.org/10.1007/s00127-017-1410-0

Bielamowicz, S., Kreiman, J., Gerratt, B. R., Dauer, M. S., & Berke, G. S. (1996). Comparison of Voice Analysis Systems for Perturbation Measurement. *Journal of Speech, Language, and Hearing Research*, *39*(1), 126–134. https://doi.org/10.1044/jshr.3901.126

Borrell, L. N., Elhawary, J. R., Fuentes-Afflick, E., Witonsky, J., Bhakta, N., Wu, A. H. B.,

Bibbins-Domingo, K., Rodríguez-Santana, J. R., Lenoir, M. A., Gavin, J. R., Kittles, R. A., Zaitlen, N. A., Wilkes, D. S., Powe, N. R., Ziv, E., & Burchard, E. G. (2021). Race and Genetic Ancestry in Medicine—A Time for Reckoning with Racism. *New England Journal of Medicine*, *384*(5), 474–480. https://doi.org/10.1056/NEJMms2029562

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., & Bolt, D. M. (2014). Quantitative and Descriptive Comparison of Four Acoustic Analysis Systems: Vowel Measurements. *Journal of Speech, Language, and Hearing Research*, *57*(1), 26–45. https://doi.org/10.1044/1092-4388(2013/12-0103)

CDC. (2022, July 7). *What is Diabetes?* Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

CDC. (2023, February 28). *Diabetes Testing*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/getting-tested.html

Chawla, A., Chawla, R., & Jaggi, S. (2016). Microvasular and macrovascular complications in diabetes mellitus: Distinct or continuum? *Indian Journal of Endocrinology and Metabolism*, *20*(4), 546–551. https://doi.org/10.4103/2230-8210.183480

Chen, X., Li, Z., Setlur, S., & Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-06673-y

Chitkara, D., & Sharma, R. K. (2016). Voice based detection of type 2 diabetes mellitus. *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 83–87. https://doi.org/10.1109/AEEICB.2016.7538402

Deutsch, D., Le, J., Shen, J., & Henthorn, T. (2009). The pitch levels of female speech in two

Chinese villages. *The Journal of the Acoustical Society of America*, *125*(5), EL208-213. https://doi.org/10.1121/1.3113892

Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers*, *5*(1), 78–88. https://doi.org/10.1159/000515346

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton & Co. N.V.

Gender and health. (2023). WHO.Int. https://www.who.int/health-topics/gender

Gibney, E. (2020). The battle for ethical AI at the world's biggest machine-learning conference. *Nature*, *577*(7792), 609–609. https://doi.org/10.1038/d41586-020-00160-y

Gugatschka, M., Kiesler, K., Obermayer-Pietsch, B., Schoekler, B., Schmid, C., Groselj-Strele, A., & Friedrich, G. (2010). Sex hormones and the elderly male voice. *Journal of Voice: Official Journal of the Voice Foundation*, *24*(3), 369–373. https://doi.org/10.1016/j.jvoice.2008.07.004

Hamdan, A., Jabbour, J., Barazi, R., Korban, Z., & Azar, S. T. (2013). Prevalence of laryngopharyngeal reflux disease in patients with diabetes mellitus. *Journal of Voice: Official Journal of the Voice Foundation*, *27*(4), 495–499. https://doi.org/10.1016/j.jvoice.2012.07.010

Hamdan, A., Jabbour, J., Nassar, J., Dahouk, I., & Azar, S. T. (2012). Vocal characteristics in patients with type 2 diabetes mellitus. *European Archives of Oto-Rhino-Laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, *269*(5), 1489–1495. https://doi.org/10.1007/s00405-012-1933-7

Hari Kumar, K. V. S., Garg, A., Ajai Chandra, N. S., Singh, S. P., & Datta, R. (2016). Voice and

endocrinology. *Indian Journal of Endocrinology and Metabolism*, *20*(5), 590–594. https://doi.org/10.4103/2230-8210.190523

Heller, A., & Feldman, B. (2008). Electrochemical Glucose Sensors and Their Applications in Diabetes Management. *Chemical Reviews*, *108*(7), 2482–2505. https://doi.org/10.1021/cr068069y

Huang, H., Hu, W., Han, Z., & Ye, H. (2005). Hybrid Progressive Algorithm to Recognize Type II Diabetic Based on Hair Mineral Element Contents. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2005*, 4716–4718. https://doi.org/10.1109/IEMBS.2005.1615524

IDF. (2023, January 16). *IDF Diabetes Atlas | Tenth Edition*. https://diabetesatlas.org/

Johnson, W., & Pauwels, E. (n.d.). *How to Optimize Human Biology:*

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics*, *22*(1), 14. https://doi.org/10.1186/s12910-021-00577-8

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY*, *35*(4), 927–936. https://doi.org/10.1007/s00146-020-00965-5

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, *322*(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058

Pinyopodjanard, S., Suppakitjanusant, P., Lomprew, P., Kasemkosin, N., Chailurkit, L., & Ongphiphadhanakul, B. (2021). Instrumental Acoustic Voice Characteristics in Adults

with Type 2 Diabetes. *Journal of Voice: Official Journal of the Voice Foundation*, *35*(1), 116–121. https://doi.org/10.1016/j.jvoice.2019.07.003

Poddar, A., Sahidullah, M., & Saha, G. (2018). Speaker verification with short utterances: A review of challenges, trends and opportunities. *IET Biometrics*, *7*(2), 91–101. https://doi.org/10.1049/iet-bmt.2017.0065

Podesva, R. J., & Callier, P. (2015). Voice Quality and Identity. *Annual Review of Applied Linguistics*, *35*, 173–194. https://doi.org/10.1017/S0267190514000270

Race. (2022, September 14). Genome.Gov. https://www.genome.gov/genetics-glossary/Race

Ravi, R., & Gunjawate, D. R. (2019). Effect of diabetes mellitus on voice: A systematic review. *Practical Diabetes*, *36*(5), 177–180. https://doi.org/10.1002/pdi.2240

Read, C., Buder, E. H., & Kent, R. D. (1992). Speech Analysis Systems. *Journal of Speech, Language, and Hearing Research*, *35*(2), 314–332. https://doi.org/10.1044/jshr.3502.314

Saghiri, M. A., Vakhnovetsky, A., & Vakhnovetsky, J. (2022). Scoping review of the relationship between diabetes and voice quality. *Diabetes Research and Clinical Practice*, *185*, 109782. https://doi.org/10.1016/j.diabres.2022.109782

Skladnev, V. N., Ghevondian, N., Tarnavskii, S., Paramalingam, N., & Jones, T. W. (2010). Clinical evaluation of a noninvasive alarm system for nocturnal hypoglycemia. *Journal of Diabetes Science and Technology*, *4*(1), 67–74. https://doi.org/10.1177/193229681000400109

Smith, J. L. (2022). *The Pursuit of Noninvasive Glucose: "Hunting the Deceitful Turkey"* (Eighth Edition: Revised and Expanded).

Statistics About Diabetes | ADA. (2023). https://diabetes.org/about-us/statistics/about-diabetes

The Fourth Industrial Revolution: What it means and how to respond. (2016, January 14). World

Economic Forum. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Weinstein, S., Suppakitjanusant, P., & Tangpricha, V. (2022). Detection of cystic fibrosis related diabetes through analysis of voice characteristics in telemedicine clinics. *Journal of Investigative Medicine*, 703–704.

Xue, S. A., & Hao, J. G. (2006). Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *Journal of Voice: Official Journal of the Voice Foundation*, *20*(3), 391–400. https://doi.org/10.1016/j.jvoice.2005.05.001

Zhang, B., Vijaya Kumar, B. V. K., & Zhang, D. (2014). Noninvasive diabetes mellitus detection using facial block color with a sparse representation classifier. *IEEE Transactions on Bio-Medical Engineering*, *61*(4), 1027–1033. https://doi.org/10.1109/TBME.2013.2292936

Zhang, X., Pérez-Stable, E. J., Bourne, P. E., Peprah, E., Duru, O. K., Breen, N., Berrigan, D., Wood, F., Jackson, J. S., Wong, D. W. S., & Denny, J. (2017). Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century. *Ethnicity & Disease*, *27*(2), Article 2. https://doi.org/10.18865/ed.27.2.95

Zhang, Z. (2016). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, *140*(4), 2614. https://doi.org/10.1121/1.4964509